

A Survey of Test-Time Compute: From Intuitive Inference to Deliberate Reasoning

YIXIN JI, Soochow University, China

JUNTAO LI, Soochow University, China

YANG XIANG, Soochow University, China

HAI YE, National University of Singapore, Singapore

KAIXIN WU, Ant Group, China

KAI YAO, Ant Group, China

JIA XU, Ant Group, China

LINJIAN MO, Ant Group, China

MIN ZHANG, Soochow University, China

The remarkable performance of the o1 model in complex reasoning demonstrates that test-time compute scaling can further unlock the model's potential, enabling powerful System-2 thinking. However, there is still a lack of comprehensive surveys for test-time compute scaling. We trace the concept of test-time compute back to System-1 models. In System-1 models, test-time compute addresses distribution shifts and improves robustness and generalization through parameter updating, input modification, representation editing, and output calibration. In System-2 models, it enhances the model's reasoning ability to solve complex problems through repeated sampling, self-correction, and tree search. We organize this survey according to the trend of System-1 to System-2 thinking, highlighting the key role of test-time compute in the transition from System-1 models to weak System-2 models, and then to strong System-2 models. We also point out advanced topics and future directions.¹

CCS Concepts: • Computer methodologies → Artificial intelligence; Nature language processing; Nature language generation.

Additional Key Words and Phrases: Large Language Models, Test-time compute, Reasoning

ACM Reference Format:

Yixin Ji, Juntao Li, Yang Xiang, Hai Ye, Kaixin Wu, Kai Yao, Jia Xu, Linjian Mo, and Min Zhang. 2025. A Survey of Test-Time Compute: From Intuitive Inference to Deliberate Reasoning. 1, 1 (July 2025), 36 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

¹https://github.com/Dereck0602/Awesome_Test_Time_LLMs.

Authors' Contact Information: Yixin Ji, jiyixin169@gmail.com, Soochow University, Suzhou, China; Juntao Li, Soochow University, Suzhou, China, ljt@suda.edu.cn; Yang Xiang, Soochow University, Suzhou, China; Hai Ye, National University of Singapore, Singapore; Kaixin Wu, Ant Group, Hangzhou, China; Kai Yao, Ant Group, Hangzhou, China; Jia Xu, Ant Group, Hangzhou, China; Linjian Mo, Ant Group, Hangzhou, China; Min Zhang, Soochow University, Suzhou, China, minzhang@suda.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

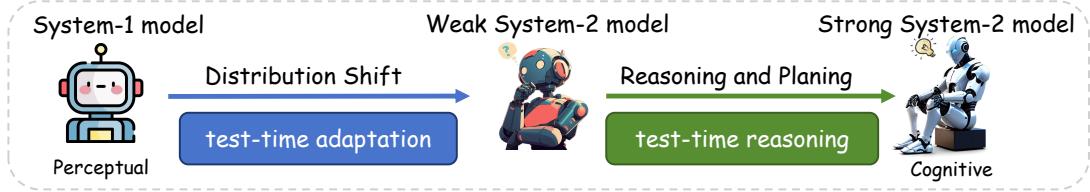


Fig. 1. Illustration of test-time compute in the System-1 and System-2 model.

1 Introduction

Over the past decades, deep learning with its scaling effects has been the driving engine behind the artificial intelligence revolution. Particularly in the text modality, large language models (LLMs) represented by the GPT series [24, 226, 229] have demonstrated that larger models and more training data lead to better performance on downstream tasks. However, on the one hand, further scaling in the training phase becomes difficult due to the scarcity of data and computational resources [299]; on the other hand, existing models still perform far below expectations in terms of robustness and handling complex tasks. These shortcomings are attributed to the model’s reliance on fast, intuitive System-1 thinking, rather than slow, deep System-2 thinking [328]. Recently, large reasoning models (LRMs), represented by OpenAI-o1/o3 [227], DeepSeek-R1 [95], and Gemini 2.5 [57], equipped with System-2 thinking, have gained attention for their outstanding performance in complex reasoning tasks. It demonstrates a test-time compute scaling effect: the greater the computational effort in the inference, the better the model’s performance.

The concept of test-time compute emerged before the rise of LLMs and was initially applied to System-1 models (illustrated in Figure 1). These System-1 models can only perform limited perceptual tasks, relying on patterns learned during training for predictions. As a result, they are constrained by the assumption that training and testing are identically distributed and lack robustness and generalization to distribution shifts [423]. Many works have explored test-time adaptation (TTA) to improve model robustness by updating parameters [302, 363], modifying the input [61], editing representations [247], and calibrating the output [402]. With TTA, the System-1 model slows down its thinking process and has better generalization. However, TTA is an implicit slow thinking, unable to exhibit explicit, logical thinking process like humans, and struggles to handle complex reasoning tasks. Thus, TTA-enabled models perform weak System-2 thinking.

Currently, advanced LLMs with chain-of-thought (CoT) prompting [326] have enabled language models to perform explicit System-2 thinking [97]. However, vanilla CoT is limited by error accumulation and linear thinking pattern [271], making it difficult to fully simulate non-linear human cognitive processes such as brainstorming, reflection, and backtracking. To achieve stronger System-2 models, researchers employ test-time compute strategies to extend model reasoning’s breadth, depth and accuracy, such as repeated sampling [50], self-correction [263], and tree search [359]. Repeated sampling simulates the diversity of human thinking, self-correction enables LLMs to reflect, and tree search enhances reasoning depth and backtracking.

To the best of our knowledge, this paper is the first to systematically review test-time compute methods and thoroughly explore their critical role in advancing models from System-1 to weak System-2, and ultimately to strong System-2 thinking. In Section 2, we present the background of System-1 and System-2 thinking. Section 3 and Section 4 detail the test-time compute methods for the System-1 and System-2 models. Then, we discuss advanced topics and future directions in Section 5 and 6. Additionally, we review benchmarks in Section 7.

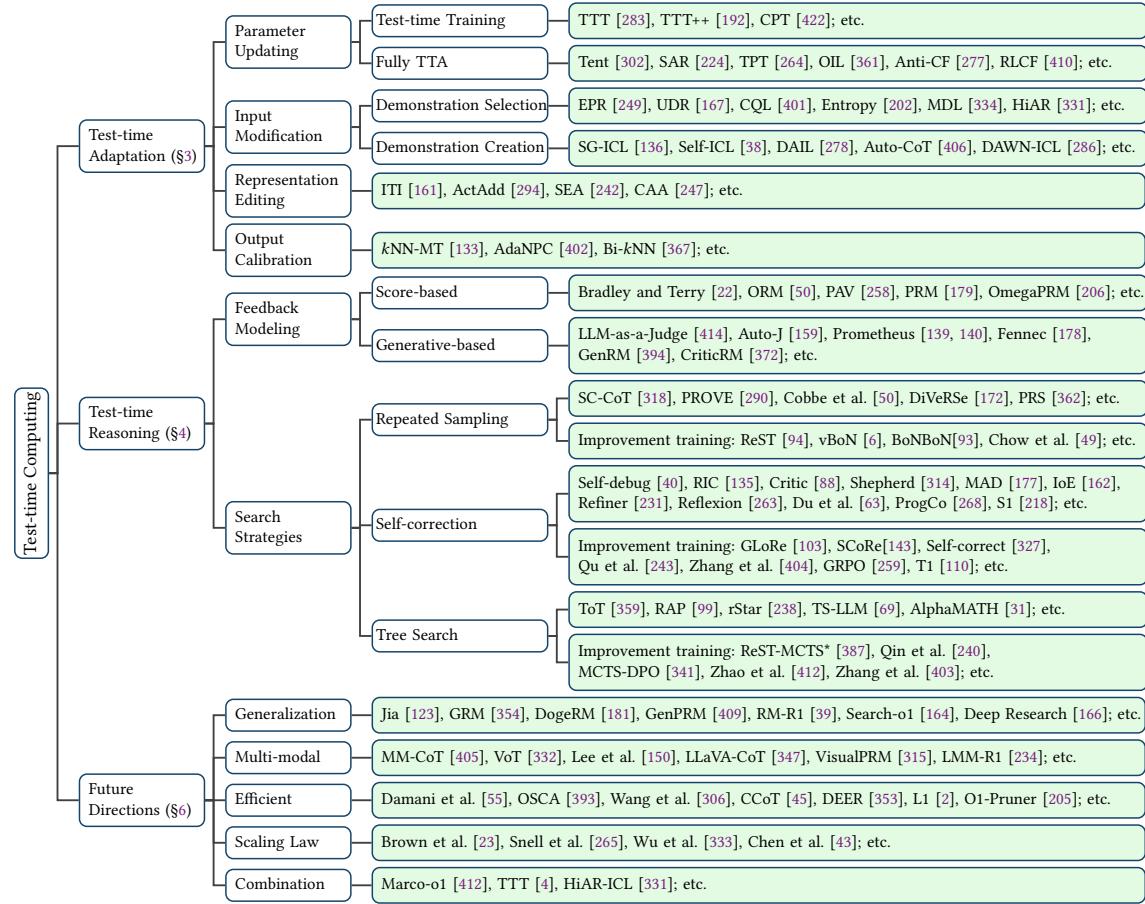


Fig. 2. Taxonomy of test-time computing methods and future directions.

2 Background

System-1 and System-2 thinking are psychological concepts [127]. When recognizing familiar patterns or handling simple problems, humans often respond intuitively. This automatic, fast thinking is called System-1 thinking. In contrast, when dealing with complex problems like mathematical proofs or logical reasoning, deep and deliberate thought is required, referred as System-2 thinking—slow and reflective. In the field of artificial intelligence, researchers also use these terms to describe different types of models [148]. System-1 models respond directly based on internally encoded perceptual information and world knowledge without showing any intermediate decision-making process. In contrast, System-2 models explicitly generate reasoning processes and solve tasks incrementally. Before the rise of LLMs, System-1 models were the mainstream in AI. Although many deep learning models, such as ResNet, Transformer, and BERT, achieve excellent performance in various tasks in computer vision and natural language processing, these System-1 models, similar to human intuition, lack sufficient robustness and are prone to errors [62, 82, 317]. Nowadays, the strong generation and reasoning capabilities of LLMs make it possible to build System-2 models. Wei et al. [326] propose the CoT, which allows LLMs to generate intermediate reasoning steps progressively during inference. Empirical and theoretical

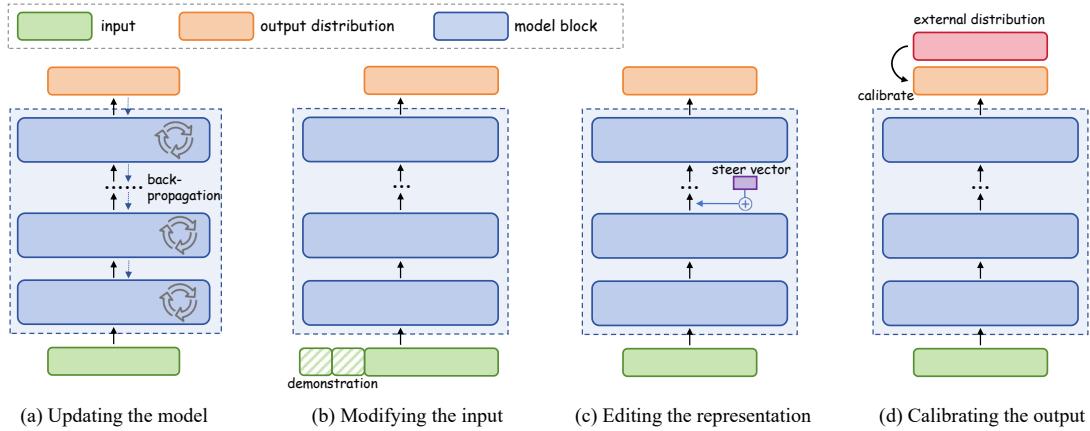


Fig. 3. Illustration of various kinds of test-time adaptation methods.

results show that this approach significantly outperforms methods that generate answers directly [67, 142, 174, 416]. However, current System-2 models represented by CoT prompting still have shortcomings. The intermediate processes generated by LLMs may contain errors, leading to cumulative mistakes and ultimately resulting in incorrect answers. Although retrieval-augmented generation (RAG) helps mitigate factual errors [91, 292, 324], their impact on improving reasoning abilities remains limited. As a result, CoT-enabled LLMs are still at the weak system-2 thinking stage.

3 Test-time Adaptation for System-1 Thinking

3.1 Updating the Model

Model updating utilizes test sample information to finetune model parameters during the inference stage, enabling the model to adapt to the test distribution. The key lies in obtaining test distribution information to provide learning signals and using appropriate parameters and optimization algorithms to achieve efficient and stable updates.

Learning signal. In the inference stage, the ground-truth of test samples is unavailable. Thus, many works attempt to design unsupervised or self-supervised objectives as learning signals. Existing learning signals can be classified into two categories based on whether the training process can be modified: *test-time training* (TTT) and *fully test-time adaptation* (FTTA). TTT assumes users can modify the training process by incorporating distribution-shift-aware auxiliary tasks. During test-time adaptation, the auxiliary task loss serves as the learning signal for optimization. Many self-supervised tasks have been shown to be effective as auxiliary tasks in image modality, such as rotation prediction [283], meta learning [18], masked autoencoding [74] and contrastive learning [30, 192]. Among them, contrast learning has been successfully applied to test-time adaptation for visual-language tasks due to its generalization of self-supervised learning within and across modalities [422].

In contrast, FTTA is free from accessing the training process and instead uses internal or external feedback on test samples as learning signals. Uncertainty is the most commonly learned signal, driven by the motivation that when test samples shift from the training distribution, the model's confidence in its predictions is lower, resulting in higher uncertainty. Tent [302] uses the entropy of model predictions as a measure of uncertainty and updates the model by minimizing the entropy. MEMO [395] augments the data for a single test sample and then minimizes its

marginal entropy, which is more stable compared to Tent in the single-sample TTA setting. However, minimizing entropy also has pitfalls, as blindly reducing prediction uncertainty may cause the model to collapse and make trivial predictions [235, 277, 408]. Some works propose new regularization terms for minimizing entropy to avoid model collapse, including Kullback-Leibler divergence [277], moment matching [102] and entropy matching [17]. For specific tasks, a small amount of human feedback or external model rewards can also serve as high-quality learning signals. Gao et al. [76] and Li et al. [175] utilize user feedback to adapt the QA model. Zhan et al. [384] apply test-time adaptation to multilingual machine translation tasks by using COMET [246] for evaluating translation quality. In cross-modal tasks such as image-text retrieval and image captioning, RLCF [410] demonstrates its effectiveness by using CLIP scores [244] as TTA signals. In language modeling, training with relevant contextual text at test time can reduce perplexity [101, 321]. Hübötter et al. [121] theoretically shows that it reduces the uncertainty of test samples and proposes a more effective active learning selection strategy.

Updating parameters. To advance the application of TTA in real-world scenarios, researchers must address challenges of efficiency and stability. To improve efficiency, many methods only fine-tune a small subset of parameters, such as normalization layers [255], soft prompt [66, 102, 151, 212, 223, 264], low-rank module [112, 113, 122], adapter module [111, 219, 277] and cross-modality projector [410]. Although the number of parameters to fine-tune is reduced, TTA still requires an additional backward propagation. Typically, the time cost of a backward propagation is approximately twice that of a forward propagation. Thus, Niu et al. [223] propose FOA, which is free from backward propagation by adapting soft prompt through covariance matrix adaptation evolution strategy.

The stability of TTA is primarily shown in two aspects. On the one hand, unsupervised or self-supervised learning signals inevitably introduce noise into the optimization process, resulting in TTA optimizing the model in the incorrect gradient direction. To address this, Niu et al. [224] and Gong et al. [86] propose noise data filtering strategies and the robust sharpness-aware optimizer. On the other hand, in real-world scenarios, the distribution of test samples may continually shift, but continual TTA optimization may lead to catastrophic forgetting of the model’s original knowledge. Episodic TTA [264, 410] is a setting to avoid forgetting, which resets the model parameters to their original state after TTA on a single test sample. However, episodic TTA frequently loads the original model, leading to higher inference latency and also limiting the model’s incremental learning capability. To overcome the dilemma, a common trick is the exponential moving average [329, 361], which incorporates information from previous model states.

3.2 Modifying the Input

When it comes to LLM, the large number of parameters makes model update-based TTA methods face a tougher dilemma of efficiency and stability. As a result, input-modification-based methods, which do not rely on parameter updates, have become the mainstream method for TTA in LLMs. The effectiveness of input-modified TTA stems from the in-context learning (ICL) capability of LLM, which can significantly improve the performance by adding some demonstrations before the test sample. ICL is highly sensitive to the selection and order of demonstrations. Therefore, the core objective of input-modification TTA is to select appropriate demonstrations for the test samples and arrange them in the optimal order to maximize the effectiveness of ICL.

First, empirical studies [186] show that the more similar the demonstrations are to the test sample, the better the ICL performance. Therefore, retrieval models like BM25 and SentenceBERT are used to retrieve demonstrations semantically closest to the test sample and rank them in descending order of similarity [207, 239]. To improve the accuracy of demonstration retrieval, Rubin et al. [249] and Li et al. [167] specifically train the demonstration retriever by contrastive

learning. Then, as researchers delve deeper into the mechanisms of ICL, ICL is considered to conduct implicit gradient descent on the demonstrations [53]. Therefore, from the perspective of training data, demonstrations also need to be informative and diverse [168, 276]. Wang et al. [319] view language models as topic models and formulate the demonstration selection problem as solving a Bayesian optimal classifier. Additionally, the ordering of examples is another important area for improvement. Lu et al. [202] and Wu et al. [334] use information theory as a guide to select the examples with maximum local entropy and minimum description length for ranking, respectively. Scarlatos and Lan [254] and Zhang et al. [401] consider the sequential dependency among demonstrations, and model it as a sequential decision problem and optimize demonstration selection and ordering through reinforcement learning.

Another line of work [38, 136, 209, 406] argues that in practice, combining a limited set of externally provided examples may not always be the optimal choice. LLMs can leverage their generative and annotation capabilities to create better demonstrations. DAIL [278] constructs a demonstration memory, storing previous test samples and their predictions as candidate demonstrations for subsequent samples. DAWN-ICL [286] further models the traversal order of test samples as a planning task and optimizes it by the Monte Carlo tree search (MCTS).

3.3 Editing the Representation

For generative LLMs, some works have found that the performance bottleneck is not in encoding world knowledge, but in the large gap between the information in intermediate layers and the output. During the inference phase, editing the representation can help externalize the intermediate knowledge into the output. PPLM [56] performs gradient-based representation editing under the guidance of a small language model to control the style of outputs. ActAdd [294] selects two semantically contrastive prompts and calculates the difference between their representations as a steering vector, which is then added to the residual stream. Representation editing based on contrastive prompts has demonstrated its effectiveness in broader scenarios, including instruction following [274], alleviating hallucinations [9, 161], reducing toxicity [188, 200] and personality [26]. SEA [242] projects representations onto directions with maximum covariance with positive prompts and minimum covariance with negative prompts. They also introduce nonlinear feature transformations, allowing representation editing to go beyond linearly separable representations. Scalena et al. [253] conduct an in-depth study on the selection of steering intensity. They find that applying a gradually decreasing steering intensity to each output token can improve control over the generation without compromising quality.

3.4 Calibrating the Output

Using external information to calibrate the model’s output distribution is also an efficient yet effective test-time adaptation method [134]. AdaNPC [402] designs a memory pool to store training data. During inference, given a test sample, AdaNPC recalls k samples from the memory pool and uses a k NN classifier to predict the test sample. It then stores the test sample and its predicted label in the memory pool. Over time, the sample distribution in the memory pool gradually aligns with the test distribution. In NLP, the most representative application of such methods is k NN machine translation (k NN-MT). k NN-MT [133] constructs a datastore to store contextual representations and their corresponding target tokens. During translation inference, it retrieves the k -nearest candidate tokens from the datastore based on the decoded context and processes them into probabilities. Finally, it calibrates the translation model’s probability distribution by performing a weighted fusion of the model’s probabilities and the retrieved probabilities. k NN-MT has demonstrated superior transferability and generalization compared to traditional models in cross-domain and multilingual MT tasks. Subsequent studies have focused on improving its performance and efficiency [301, 367, 421] or applying its methods to other NLP tasks [20, 312].

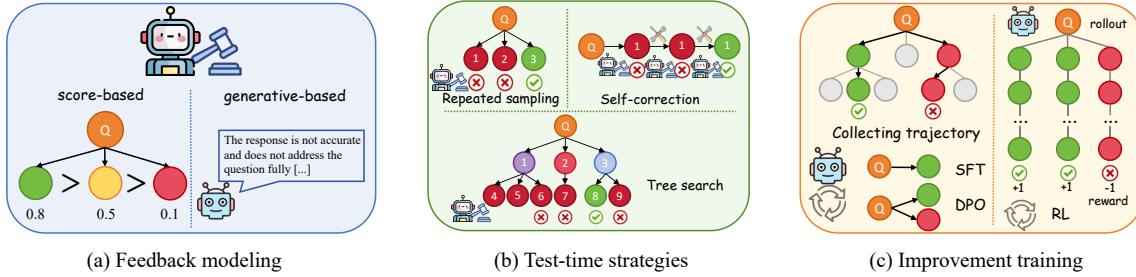


Fig. 4. Illustration of feedback modeling, search strategies and improvement training in test-time reasoning.

Summary 1: Parameter updating and output calibration are the most versatile TTA methods. However, parameter updating suffers from training instability and inefficiency in LLMs, while output calibration relies on target domain information and risks knowledge leakage. Input modification and representation editing are free from training but have limited applicability: input modification is related to ICL capabilities, and representation editing demands manual prior knowledge.

4 Test-time Reasoning for System-2 Thinking

Test-time reasoning aims to spend more inference time to search for the most human-like reasoning process within the vast decoding search space. In this section, we introduce the two core components of test-time reasoning: feedback modeling and search strategies (as shown in Figure 4).

4.1 Feedback Modeling

Score-based Feedback. Score-based feedback, also known as the verifier, aims to predict scalar scores to evaluate the alignment of generated results with ground truth or human cognitive processes. Its training process is typically similar to the reward model in RLHF, using various forms of feedback signals and modeling it as a classification [50] or rank task [22, 108, 375]. In reasoning tasks, verifiers are primarily categorized into two types: outcome-based verifiers (ORMs) and process-based verifiers (PRMs). ORMs [50] use the correctness of the final CoT result as training signals. Liu et al. [196] provide a detailed recipe for training a strong ORM. In contrast, PRMs [179, 296, 390] are trained based on the correctness of each reasoning step. Compared to ORMs, PRMs cannot only evaluate intermediate reasoning steps but also assess the entire reasoning process more precisely. However, PRMs require more human effort to annotate feedback for the intermediate steps. Math-Shepherd [311], OmegaPRM [206] and EpicPRM [282] utilize MCTS algorithm to collect high-quality process supervision data automatically. Zhang et al. [407] utilize critic models to evaluate process annotations collected by MCTS, filtering out low-quality data to improve the training effectiveness of PRMs. Setlur et al. [258] argue that PRMs should evaluate the advantage of each step for subsequent reasoning rather than focusing solely on its correctness. They propose process advantage verifiers (PAVs) and efficiently construct training data through MCTS. Furthermore, Lu et al. [201] and Yuan et al. [376] notice that ORMs implicitly model the advantage of each step, leading them to automatically annotate process supervision data using ORMs or directly train PRMs on outcome labels, respectively. Liu et al. [193] adaptively segment reasoning steps based on tokens with higher uncertainty, which is used to train better PRMs.

Category	Sub-category	Representative Methods	Domain	Objective	Description
Score-based	ORM	Cobbe et al. [50] Acemath [196]	Math	classification list-wise Bradley-Terry	ORM; human annotated data ORM; sampling training data from multiple LLMs
		Lightman et al. [179] Math-shepherd [311]	Math	classification	PRM; human annotated data
	PRM	Zhang et al. [407]	Math	classification	PRM; annotating processes via MC estimation
		Implicit PRM [376]	Math	classification/regression	PRM; annotating processes via MCTS and LLM-as-a-judge
		ASPRM [193]	Math, Code	implicit reward modeling	PRM; training PRMs with outcome labels
				classification	PRM; adaptive segment step; annotating processes via MC estimation
Generative-based	Training-free	LLM-as-a-Judge [414]	General	-	Designing system instructions to mitigate biases
		BSM [250]	General	-	Dividing into multiple criteria and then merging
	Training-based	Shepherd [314]	General	SFT	Collecting data from human annotation and the Internet
		Prometheus [140]	General	SFT	Training single and pairwise models and then merging them
		EvalPlanner [251]	General	DPO	Planning evaluation processes and then evaluating
		GenRM [394]	Math	SFT	PRM; synthesizing critique data from external LLMs
		R-PRM [260]	Math	SFT & DPO	PRM; synthesizing critique data from external LLMs
		Critic-RM [372]	General	SFT & Bradley-Terry	ORM; synthesizing and filtering critique data via self-critique
		Cloud [8]	General	SFT & Bradley-Terry	ORM; synthesizing data from external LLMs and self-critique

Table 1. Overview of feedback modeling methods.

Generative-based Feedback. Although the verifier can evaluate the correctness of generated answers or steps, it lacks interpretability, making it unable to locate the specific cause of errors or provide correction suggestions. Generative-based feedback, also referred to critic, fully leverages the LLM’s generative and instruction-following ability [8, 48, 260, 364]. By designing specific instructions, it can perform pointwise or pairwise evaluation from multiple dimensions, and even provide suggestions for revision in natural language. Powerful closed-source LLMs, such as GPT-4 and Claude, are effective critics. They can perform detailed and controlled assessments of generated texts, such as factuality, logical errors, coherence, and alignment, with high consistency with human evaluations [47, 191, 208, 308]. However, they still face biases such as length, position, and perplexity [19, 275, 310]. LLM-as-a-Judge [414] carefully designs system instructions to mitigate the interference of biases. BSM [250] evaluates based on multiple criteria and then merges them. Peng et al. [233] employ multi-agents to jointly evaluate answers’ factuality and instruction-following.

To obtain cheaper verbal-based feedback, open-source LLMs can also serve as competitive alternatives through supervised fine-tuning (SFT) [178, 231, 323, 420]. Shepherd [314] collects high-quality training data from human annotation and online communities to fine-tune an evaluation model. Auto-J [159] collects queries and responses from various scenarios and designs evaluation criteria for each scenario. GPT-4 then generates critiques of the responses based on these criteria and distills its critique ability to open-source LLMs. Prometheus [139, 140] designs more fine-grained evaluation dimensions. It trains a single evaluation model and a pairwise ranking model separately, then unifies them into one LLM by weight merging. To reduce reliance on human annotations and external LLMs, Wang et al. [313] propose a self-training method: the critique model generates positive and negative responses, then collects critique data via rejection sampling to perform iterative finetuning. Building on self-training, EvalPlanner enables [251] the critique model to plan evaluation processes and criteria, conduct critiques based on these, and then collect positive and negative samples to improve the critique model via DPO [245]. [370] carefully synthesize and filter data, enabling the base model to achieve strong critique abilities with only 40K samples for SFT and DPO training. GenRM [394] leverages instruction tuning to enable the verifier to answer ‘Is the answer correct (Yes/No)?’ and uses the probability of generated ‘Yes’ token as the score. GenRM can also incorporate CoT, allowing the verifier to generate the corresponding rationale before answering ‘Yes’ or ‘No’. ThinkPRM [131] evaluates each reasoning step with long CoT and requires only 1% of the process supervision data compared to discriminative PRMs. Critic-RM [372] jointly trains the critic and the verifier.

Category	sub-category	Representative Methods	Tasks	Verifier/Critic	Train-free
Repeat Sampling	Majority voting	CoT-SC [318] PROVE [290]	Math, QA Math	self-consistency compiler	✓ ✓
	Best-of-N	Cobbe et al. [50]	Math	ORM	✗
		DiVeRSe [172]	Math	PRM	✗
		Knockout [195]	Math	critic	✓
Self-correction	Human feedback	NL-EDIT [64] FBNET [285]	Semantic parsing Code	Human Human	✗ ✗
	External tools	DrRepair [360]	Code	compiler	✗
		Self-debug [40]	Code	compiler	✓
		CRITIC [88]	Math, QA, Detoxifying	text-to-text APIs	✓
	External models	REFINER [231]	Math, Reason	critic model	✗
		Shepherd [314]	QA	critic model	✗
		Multiagent Debate [63]	Math, Reason	multi-agent debate	✓
Intrinsic feedback	MAD [177]	MAD [177]	Translation, Math	multi-agent debate	✓
		Self-Refine [214]	Math, Code, Controlled generation	self-critique	✓
		Reflexion [263]	QA	self-critique	✓
	RCI [135]	RCI [135]	Code, QA	self-critique	✓
Tree Search	Uninformed search	ToT [359] Xie et al. [342]	Planing, Creative writing Math	self-critique self-critique	✓ ✓
	Heuristic search	RAP [99]	Planing, Math, Logical	self-critique	✓
		TS-LLM [69]	Planing, Math, Logical	ORM	✗
		rStar [238]	Math, QA	multi-agent consistency	✓
	ReST-MCTS* [387]	ReST-MCTS* [387]	Math, QA	PRM	✗

Table 2. Overview of search strategies.

4.2 Search Strategies

4.2.1 Repeated Sampling.

Sampling strategies such as top-p and top-k are commonly used decoding algorithms in LLM inference. They introduce randomness during decoding to enhance text diversity, allowing for parallelly sampling multiple generated texts. Through repeated sampling, we have more opportunities to find the correct answer. Repeated sampling is particularly suitable for tasks that can be automatically verified, such as code generation, where we can easily identify the correct solution from multiple samples using unit tests [169, 248]. For tasks that are difficult to verify, like math word problems, the key to the effectiveness of repeated sampling is the verification strategy.

Verification strategy. Verification strategies include two types: majority voting and best-of-N (BoN) sampling. *Majority voting* [160, 180] selects the most frequently occurring answer in the samples as the final answer, which is motivated by ensemble learning. Majority voting is simple yet effective. For instance, self-consistency CoT [318] can improve accuracy by 18% over vanilla CoT in math reasoning tasks. However, the majority does not always hold the truth, as they may make similar mistakes. Therefore, some studies perform filtering or weighting before voting. For example, the PROVE framework [290] converts CoT into executable programs, filtering out samples if the program’s results are inconsistent with the reasoning chain’s outcomes. Kang et al. [130] propose the self-certainty metric to weight votes based on their ranking. Huang et al. [116] use the calibrated confidence as vote weights, and adjust the number of samples based on response agreement to improve efficiency [1]. RPC [419] demonstrates theoretically and experimentally that filtering low-probability samples improves the performance and efficiency of majority voting.

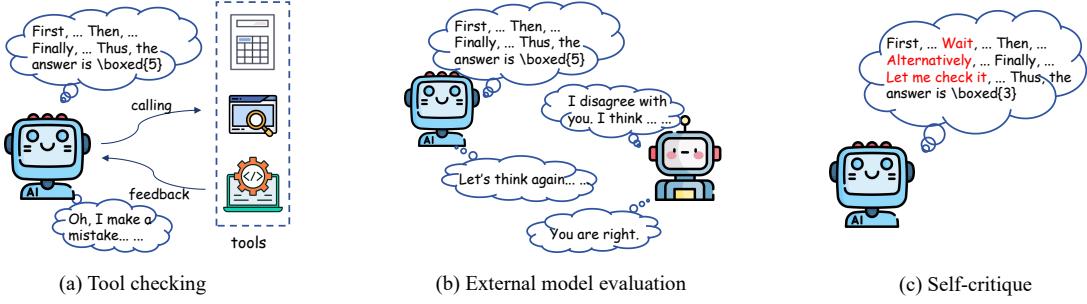


Fig. 5. Illustration of self-correction with tool checking, external model evaluation, and self-critique.

Best-of-N sampling uses a verifier to score each response and selects the one with the highest score as the final answer [50, 184, 221, 273]. Li et al. [172] propose a voting-based BoN variant, which performs weighted voting on all answers based on the verifier’s scores and selects the answer with the highest score. [195] design BoN in a knockout tournament, using pairwise comparison verifiers to filter out the best response. In addition, some works aim to improve the efficiency of BoN. Inspired by speculative decoding, Qiu et al. [241], Sun et al. [281], Yu et al. [368], Zhang et al. [397] and Manvi et al. [215] evaluate each reasoning step and prune low-scoring sampled results, halting further generation for those paths, thereby significantly reducing the overall time cost. PRS [362] enables LLMs to self-critique and self-correct, guiding the model to generate expected responses with fewer sampling times. Li et al. [156] compare BoN and majority voting, demonstrating that BoN is suitable for harder questions with moderate diversity of response distribution.

Improvement Training. Repeated sampling has proven to be a simple yet effective method, even surpassing models fine-tuned with RLHF [80, 109]. However, it costs much inference time that is difficult to afford in practical applications. Therefore, many studies have attempted to train the model by BoN sampling to approximate the BoN distribution, thereby reducing the search space during inference. ReST [94] samples responses with reward values above a threshold from the policy model as self-training data and fine-tunes the policy model by offline reinforcement learning. In each iteration, ReST samples new training data. vBoN [6], BoNBoN [93] and BOND [257] derive the BoN distribution and minimize the difference between the policy model’s distribution and the BoN distribution. Chow et al. [49] design a BoN-aware loss to make the policy model more exploratory during fine-tuning.

4.2.2 Self-correction.

Self-correction is a sequential test-time compute method that enables LLMs to iteratively revise and refine generated results based on external or internal feedback [263].

Feedback sources. The feedback used for self-correction is typically presented in natural language and comes from various sources, including human evaluation, tool checking, external model evaluation, and intrinsic feedback. *Human evaluation* is the gold standard for feedback, but due to its high cost and limited scalability, it is mainly used in early research to explore the upper limits of self-correction capabilities [64, 284, 285]. For certain domain-specific tasks, *external tool checking* provides accurate and efficient feedback [40, 79, 88]. For example, Yasunaga and Liang [360] propose to obtain feedback from compilers in code repair and generation tasks. In embodied tasks, the environment can provide precise feedback on the action trajectories of LLM-based agents [304].

External model evaluation is an effective feedback source for general tasks, such as various verbal-based critique models described in Section 4.1. For example, Paul et al. [231] first define multiple error types for natural language reasoning tasks and then design the corresponding feedback templates. They train an evaluation model using synthetic feedback training data, and with the critic, the reasoning model achieves substantial performance improvement. Multi-agent debate [34, 63, 177, 309, 344] is another mechanism that leverages external feedback to enhance reasoning capabilities. In this approach, models do not have distinct roles as reasoners and critics. Instead, multiple models independently conduct reasoning, critique each other, and defend or refine their reasoning based on feedback. This process continues until agents reach a consensus or a judge model summarizes the final reasoning results. The multi-agent debate has shown its potential in fact-checking [132, 137], commonsense QA [344], faithful evaluations [28], and complex reasoning [63]. However, multi-agent debate may be unstable, as LLMs are susceptible to adversarial information and may revise correct answers to incorrect ones in response to misleading inputs [5, 144]. Therefore, a successful multi-agent debate requires that LLMs maintain their stance when faced with incorrect answers from other models while remaining open to valid suggestions [272]. In general, the more LLMs involved in the debate, the stronger the overall reasoning performance. However, this significantly increases the number of LLM inferences required, and the length of input context, posing a major challenge to LLM inference costs [190]. To reduce debate inference costs, Li et al. [170] investigate the impact of topological connections among multiple agents and show that sparse connections, such as ring structures, are not inferior to the fully connected topology. GroupDebate [190] divides LLMs into groups that conduct debates internally and only share the consensus results between groups.

Self-critique assumes that LLMs can self-evaluate their outputs and optimize them through intrinsic feedback [379]. This idea stems from a fundamental principle in computational complexity theory: verifying whether a solution is correct is typically easier than solving the problem. Bai et al. [16] propose self-correcting harmful responses from LLMs by prompting themselves. Self-Refine [214] and RCI Prompting [135] iteratively prompt LLMs to self-correct their responses in tasks such as arithmetic reasoning. IoE [162] observes that LLMs may over-criticize themselves during self-critique, leading to performance degradation, and designs prompt to guide LLMs in assessing confidence. ProgCo [268] leverage the advantages of code in expressing complex logic, enabling LLMs to generate responses in pseudo-code form, followed by self-critique and refinement. SETS [33] combines the strengths of repeated sampling and self-critique, applying self-critique and correction to each sampled reasoning path and choosing the final solution via majority voting. S1 [218] adds the “wait” to the reasoning process, prompting the LLM to critique its reasoning.

Arguments. The effectiveness of self-correction, especially the self-critique, has remained controversial. Several empirical studies on code generation [225], commonsense QA [117], math problem-solving [307], planning [297], and graph coloring [270] confirm that self-correction is not a guaranteed solution for improving performance. Kamoi et al. [128] think the effectiveness of self-correction has been overestimated. Previous successes either rely on oracle answers or weak initial answers. Only tasks that can be broken down into easily verifiable sub-tasks can truly benefit from self-correction. They suggest fine-tuning specific evaluation models to achieve better self-correction. Zhang et al. [396] try to interpret and alleviate the failure of self-critique via human-like cognitive bias. Tyen et al. [295] decouple the abilities of LLMs to identify and correct errors and create the corresponding evaluation datasets. The evaluation results show that LLMs do not lack the ability to correct errors during self-correction, and their main performance bottleneck lies in locating the errors. Yang et al. [357] decompose self-critique into confidence and critique capabilities. Empirical studies show that fine-tuning is necessary to enhance both capabilities simultaneously, while prompt engineering can only achieve a trade-off.

Improvement Training. Most of the aforementioned self-correction methods demonstrate significant performance improvements on advanced closed-source large models or open-source LLMs with over 70B parameters. However, for medium-scale open-source models with weaker capabilities, we need to further fine-tune them to unlock their self-correction capabilities. Supervised fine-tuning optimizes the model using high-quality multi-turn correction data, either manually annotated [252], self-rationalize [378, 381], multi-agent debate [280] or sampled from stronger LLMs [7, 78, 231, 243, 335, 404]. GLoRe [103] considers that LLMs need global or local refinement for different types of errors. To address this, they construct training sets for global and local refinement, train verifiers to identify global and local errors, and develop LLMs for refinement based on different global or local feedback signals. Xi et al. [335] design a scalable framework for synthesizing self-correction training data, enabling reasoning models to generate controlled errors and receive feedback from critics to self-correct. Although SFT is effective, training data from offline-generated self-correction trajectories can only simulate limited correction patterns. This leads to the distribution mismatch with the actual self-correction behavior during model inference. Self-correct [327] adopts online imitation learning, re-sampling new self-correction trajectories for training after each training epoch. To further expand the exploration space of LLMs, many studies adopt flexible RL algorithms to surpass the performance limits of SFT. SCoRe [143] proposes using the multi-turn RL method to improve self-critique and self-correction capability. T1 [110] employs self-correction training data for SFT cold-start, followed by RL training using the RLOO algorithm [3]. During the RL phase, high-temperature sampling and entropy rewards encourage the LLM to explore more diverse reasoning paths. Deepseek-R1 [95] uses rule-based rewards and the GRPO algorithm [259] for RL training. It also demonstrates RL’s immense potential, even without SFT cold-start, its exploration capabilities suffice to endow LLMs with strong reasoning abilities.

4.2.3 Tree Searching.

Repeated sampling and self-correction scale test-time compute in parallel and sequentially, respectively. Human thinking is a tree search that combines brainstorming in parallel with backtracking to find other paths to solutions when it encounters a dead end. Search algorithms and value functions are two critical components in tree searching.

Search algorithm. In LLM reasoning, current search algorithms include uninformed search and heuristic search. Uninformed search explores the search space according to a fixed rule. For example, tree-of-thought (ToT) [359] adopts the BFS or DFS to search, while Xie et al. [342] use beam search. Uninformed search is usually less efficient for problems with large search spaces, so heuristic search strategies represented by A* [216, 300] and MCTS [21, 99, 230] are widely used in reasoning tasks. MCTS, which eliminates the need for explicit heuristics, leverages stochastic simulations and adaptive tree expansion under uncertain environments, making it well-suited for large state spaces. It optimizes search results gradually through four steps: selection, expansion, simulation, and backpropagation, approaching the optimal solution. In contrast, A* uses a heuristic function-guided deterministic search to guarantee optimal paths, but its performance depends on the design of the heuristic function. As a result, MCTS has been successfully applied to tasks such as RAG [68, 114, 124, 158], QA [73, 203], hallucinations mitigation [46], text-to-SQL [377], etc. Additionally, Long [198] trains an LLM controller using reinforcement learning to guide the LLM reasoner’s search path, and Chari et al. [29] utilizes ant colony evolutionary algorithm to guide tree search.

Value function. The value function evaluates the value of each state and guides the tree to expand towards branches with higher values in heuristic tree search. Xu [348] train an energy function by noise-contrastive estimation as the value function. RAP [99] designs a series of heuristic value functions, including the likelihood of the action, the confidence of the state, self-evaluation results, and task-specific reward, and combines them according to task requirements. Reliable Manuscript submitted to ACM

and generalized value functions facilitate the application of MCTS to more complex problems with deeper search spaces. AlphaMath [31] and TS-LLM [69] replace the hand-crafted value function with a learned LLM value function, automatically generating reasoning process and step-level evaluation signals in MCTS. VerifierQ [236] integrates implicit Q-learning and contrastive Q-learning to train the value function, effectively mitigating the overestimation issue at the step level. Traditional MCTS methods expand only one trajectory, while rStar [238] argues that the current value function struggles to guide the selection of the optimal path accurately. Therefore, rStar retains multiple candidate paths and performs reasoning with another LLM, ultimately selecting the path where both LLMs' reasoning results are consistent. Gao et al. [81] propose SC-MCTS, which combines multiple reward models, including contrastive reward, likelihood, and self-evaluation as value functions. MCTS_r [385] and SR-MCTS [386] take complete responses as nodes, expanding the search space through self-critique and correction. SR-MCTS utilizes pairwise preference rewards and global quantile score as the value function, offering a more robust value function estimation.

Improvement Training. Tree search can guide LLMs to generate long reasoning processes, and these data help train LLMs with stronger reasoning abilities [92, 346, 383]. ReST-MCTS* [387] uses process rewards as a value function to guide MCTS, collecting high-quality reasoning trajectories and the value of each step to improve the policy model and reward model. Due to the step-by-step exploration of tree search, it can obtain finer-grained step-level feedback signals. MCTS-DPO [341] collects step-level preference data through MCTS and uses DPO for preference learning. AlphaLLM-CPL [316] ranks trajectories based on preference reward gaps and policy prediction gaps, employing curriculum learning to efficiently utilize MCTS-collected trajectories. Recently, many LMRs [240, 403, 412] have also confirmed the necessity of using tree search to construct high-quality long reasoning chain data for training.

Summary 2: *Repeated sampling is easy to implement and improves answer diversity, making it suitable for open-ended or easily verifiable tasks, though computationally inefficient. Self-correction relies on precise, fine-grained feedback and works well for easily verifiable tasks, but may not perform well with poor feedback or weak reasoning capability. Tree search optimizes complex planning tasks globally but involves complex implementation.*

5 Advanced Topics

5.1 Generalizable System-2 Model

Currently, most LMRs exhibit strong reasoning performance in specific domains such as math and code, but they struggle to generalize to cross-domain, cross-lingual, or general tasks [266, 411]. On the one hand, as the foundation of current System-2 models, CoT shows little effectiveness in non-symbolic reasoning tasks [269]. On the other hand, verifiers and critics have limited generalization capabilities, making it difficult to provide effective guidance to the reasoning model [35, 138, 154]. For the former limitation, integrating external tools and multi-agent feedback is a promising direction [147, 222]. Many works aim to enable reasoning models to learn to perform retrieval or tool use during the reasoning process. Search-o1 [164] triggers external search tools by generating special tokens that represent search actions, and Search-R1 [126] further enhances the interaction between the LLM and search tools through RL. Deep Research [87, 166, 228, 415] integrates various tools such as web search and code execution, demonstrating distinguished performance on general tasks like information synthesis, report writing, and complex code generation. For the latter challenge, some works utilize multi-objective training [305], model ensemble [181], soft reward [279] or regularization constraints [123, 354] to make verifiers more generalizable. Compared to discriminative verifiers, generative critics inherit the generalization ability of LLMs and can leverage test-time compute to realize greater

potential [141]. GenPRM [409] and GRM [197] leverage repeated sampling for test-time compute, and GRM trains a Meta RM to evaluate the quality of the critic’s outputs. RM-R1 [39] and JudgeLRM [37] use RL training to enable the critic with self-correction ability.

Additionally, weak-to-strong generalization [25] is a topic worth further exploration. People are no longer satisfied with solving mathematical problems with standard answers; they hope System-2 models can assist in scientific discovery and the proofs of mathematical conjectures. In such cases, even human experts struggle to provide accurate feedback, while weak-to-strong generalization offers a promising direction to address this issue [287].

5.2 Multimodal Test-time Compute

Visual, audio, video and other modalities are crucial for models understanding and interaction with the world. Like LLMs, large multimodal models (LMMs) also face challenges in scaling training compute. Thus, enabling efficient test-time scaling for multimodal tasks has become an advanced research topic. In System-1 thinking, TTA has been successfully applied to LMMs, improving performance in tasks such as zero-shot image classification, image-text retrieval, and image captioning [410]. For multimodal reasoning tasks, the exploration of multimodal CoT [77, 150, 217, 332, 398, 405] and multimodal critics or verifiers [293, 345] open up the possibility of building multimodal System-2 models. For example, IoT [418] and SketchPAD [115] utilize visual tools to obtain critical visual information and integrate it into CoT to refine multimodal reasoning. VisualPRM [315] constructs process annotation data through Monte Carlo estimation and trains a lightweight PRM verifier; and cutting-edge LMMs such as GPT-4o and Qwen-VL can serve as critics. Building on these foundations, repeated sampling [182] and tree search [60, 358] can effectively further improve LMMs’ multimodal reasoning performance. Xu et al. [347] divide the visual reasoning process into four stages: task summary, caption, reasoning, and answer conclusion. They propose a stage-level beam search method, which repeatedly samples at each stage and selects the best result for the next stage. Audio-Reasoner [343] transfers the four-stage CoT framework to the audio modality. To equip LMMs with the self-correction ability, Deepseek-R1’s training recipe has also been practiced in LMMs [32, 213, 261]. OpenVLThinker [58] and Vision-R1 [118] convert images into text captions, enabling text-only LRM to incorporate visual information for generating CoTs. This serves as cold-start data for further reinforcement learning to train multimodal reasoning models. LMM-R1 [234] demonstrates that the reasoning ability acquired through text-only reinforcement learning can serve as an effective cold start for multimodal reinforcement learning training.

Besides understanding and reasoning tasks, multimodal generation tasks can also benefit from test-time compute. Mainstream multimodal generative models are divided into diffusion-based [232, 356] and autoregressive-based [289, 340] models. While diffusion models lack explicit CoTs, they can expand the search space at test time by sampling various Gaussian noises [339]. For video generation, due to the high computational cost of repeatedly sampling full videos, Cong et al. [51] use PRM to verify frames individually and apply beam search to prune low-scoring frames. Liu et al. [185] propose frame tree search, using critics to evaluate videos at early, middle, and late stages with different criteria, retaining only the top-k scoring frames at each step. Autoregressive models treat intermediate generated images as reasoning steps. Guo et al. [96] propose a Potential Assessment Reward Model (PARM) to evaluate the quality potential of these intermediate images and introduce a reflection mechanism that enables the generative model to self-correct low-quality images.

5.3 Efficient Test-time Compute

The successful application of test-time compute shows that sacrificing reasoning efficiency can lead to better reasoning performance. However, researchers continue to seek a balance between performance and efficiency, aiming to achieve Manuscript submitted to ACM

optimal performance under a fixed reasoning latency budget. This requires adaptively allocating computational resources for each sample. Damani et al. [55] train a lightweight module to predict the difficulty of a question, and allocate computational resources according to its difficulty. Zhang et al. [393] further extend the allocation targets to more hyperparameters. Chen et al. [41] and Wang et al. [320] systematically evaluate the over-thinking and under-thinking phenomena in LRM, where the former leads models to overcomplicate simple problems, and the latter causes frequent switching of reasoning paths on difficult problems, thereby reducing reasoning efficiency. These phenomena even cause LRM to perform poorly when dealing with simple problems [399]. There are still many open questions worth exploring, such as how to integrate inference acceleration strategies, e.g. model compression [120, 165, 176], token pruning [71, 389], and speculative decoding [153, 337] with test-time compute, and how to allocate optimal reasoning budget according to problem difficulty [45, 98, 306].

Long CoT leads to inference latency and memory footprints of key-value cache. Recently, numerous studies have explored various strategies to reduce the reasoning length. Early work primarily focus on SFT models [52, 129, 189, 211, 220, 336, 355, 371] using variable-length CoT data. In contrast, mainstream approaches incorporate length-based rewards into reinforcement learning (RL) [11, 205, 288, 365, 366, 374] to encourage concise responses. Xiao et al. [338] measure the complexity of vision-language reasoning tasks and design a complexity-aware curriculum learning and GRPO algorithm. They organize training samples from hard to easy and dynamically adjust the reward function and KL term weights based on task complexity. L1 [2] and Elastic Reasoning [350] can enable precise control over response length based on a given token budget. Further, SelfBudgeter [173] can autonomously estimate the optimal token budget before generation, and allow users to choose whether to wait for the full response or terminate early during the generation process. However, both SFT and RL require substantial computational resources. Several studies have instead explored training-free approaches for efficient reasoning, including prompting [14, 98, 149, 349], model merging [288, 330] and early exit [72, 210, 353]. Notably, DEER [353] makes early-exit decisions during CoT generation based on the confidence of intermediate answers, effectively reducing reasoning overhead. Subsequently, Dai et al. [54] introduces a new RL paradigm that encourages the model to incentivize early thinking termination when appropriate. Another promising approach to achieving efficiency is hybrid reasoning, which dynamically switches between concise responses and long-chain reasoning based on task complexity. These methods [65, 125, 199, 204, 391] follow a similar pipeline: they start with SFT on a mixed dataset of long and short CoT samples as a cold start, and then refine the policy optimization algorithm to train the model to acquire hybrid reasoning capabilities.

6 Future Directions

6.1 Test-time Scaling Law

Unlike training-time computation scaling, test-time compute still lacks a universal scaling law. Some works have attempted to derive scaling laws for specific test-time compute strategies [152, 333]. Brown et al. [23] demonstrate that the performance has an approximately log-linear relationship with repeated sampling times. Chen et al. [43] models repeated sampling as a knockout tournament and league-style algorithm, proving theoretically that the failure probability of repeated sampling follows a power-law scaling. Snell et al. [265] investigate the scaling laws of repeated sampling and self-correction, and propose the computing-optimal scaling strategy. There are two major challenges to achieving a universal scaling law: first, current test-time compute strategies are various, each with different mechanisms to steer the model; thus, it lacks a universal framework for describing them; second, the performance of test-time

compute is affected by a variety of factors, including the difficulty of samples, the accuracy of feedback signals, and decoding hyperparameters, and we need empirical studies to filter out the critical factors.

6.2 Strategy Combination

Different test-time compute strategies are suited to various tasks and scenarios, so combining multiple strategies is one way to achieve better System-2 thinking. For example, Marco-o1 [412] combines the MCTS and self-correction, using MCTS to plan reasoning processes, and self-correction to improve the accuracy of each step. TPO [171] combines BoN sampling and self-correction. Moreover, test-time adaptation strategies in System-1 models can also be combined with test-time reasoning strategies. Akyürek et al. [4] combine test-time training with repeated sampling. They further optimize the language modeling loss on test samples, then generate multiple candidate answers through data augmentation, and finally determine the answer by majority voting. They demonstrate the potential of test-time training in reasoning tasks, surpassing the human average on the ARC challenge. Therefore, we think that for LLM reasoning, it is crucial to focus not only on emerging test-time strategies but also on test-time adaptation methods. By effectively combining these strategies, we can develop System-2 models that achieve or surpass o1-level performance.

6.3 New Test-time Compute Paradigms

Latent CoT reasoning has emerged as a promising paradigm for test-time compute. By performing internal reasoning in latent space, it overcomes the limitations of conventional explicit CoT in terms of computational efficiency and abstract reasoning capability [89, 262, 351, 352]. Coconut [100] performs reasoning in a continuous latent space by directly feeding the model’s final hidden state as the next-step input embedding, thereby giving rise to advanced reasoning patterns. CCoT [45] introduces a framework that generates continuous and variable-length contemplation tokens as compressed representations of reasoning chains, enabling efficient and accurate reasoning. LightThinker [392] improves LLM efficiency by compressing intermediate reasoning steps into compact gist tokens via specialized data construction and attention mask design, enabling the model to discard verbose chains while maintaining the ability to reason over condensed representations. Latent CoT currently faces several challenges, including the difficulty of direct supervision, limited generalization, and concerns over interpretability [42]. This calls for further exploration of token-level strategies and internal mechanisms to enable more robust and transparent latent reasoning.

7 Benchmarks

Test-time Adaptation. In System-1 models, distribution shifts include adversarial robustness, cross-domain and cross-lingual scenarios. In the field of CV, ImageNet-C [107], ImageNet-R [105], ImageNet-Sketch [303] are common datasets for TTA. Yu et al. [373] propose a benchmark to conduct a unified evaluation of TTA methods across different TTA settings and backbones on 5 image classification datasets. For NLP tasks, TTA is primarily applied in QA and machine translation tasks, with commonly used datasets such as MLQA [155], XQuAD [12], MRQA [70] and CCMATRIX [256].

Feedback Modeling. RewardBench [146] collects 20.2k prompt-choice-rejection triplets covering tasks such as dialogue, reasoning, and safety. It evaluates the accuracy of reward models in distinguishing between chosen and rejected responses. RM-Bench [194] further evaluates the impact of response style on reward models. RMB [417] extends the evaluation to the more practical BoN setting, where reward models are required to select the best response from multiple candidates. CriticBench [183] is specifically designed to evaluate a critic model’s generation, critique, and correction capabilities. For PRM, Song et al. [267] propose PRMBench, which evaluates PRMs whether they can identify the earliest incorrect

reasoning step in math tasks. ProcessBench [413] provides a more fine-grained evaluation, including redundancy, soundness, and sensitivity. In addition, there are benchmarks for evaluating multimodal feedback modeling, such as VL-RewardBench [163], VCR-Bench [237] and MJ-Bench [44].

Test-time Reasoning. Reasoning capability is the core of System-2 models, including mathematics, code, commonsense, planning, etc [382]. *Math reasoning* is one of the most compelling reasoning tasks. With the advancements in LLM and test-time compute, the accuracy on some previously challenging benchmarks, like GSM8K [50] and MATH [106], have surpassed the 90% mark. Thus, more difficult college admissions exam [10, 15, 400] and competition-level [75] math benchmarks have been proposed. Some competition-level benchmarks are not limited to textual modalities in algebra, logic reasoning, and word problems. For instance, OlympiadBench [104], OlympicArena [119] and AIME [380] provide images for geometry problems, incorporating visual information to aid in problem-solving, while AlphaGeometry [291] employs symbolic rules for geometric proofs. The most challenging benchmark currently is FrontierMath [84], with problems crafted by mathematicians and covering major branches of modern mathematics. Even the most advanced o3 has not achieved 30% accuracy.

Code ability is a key aspect of LLM reasoning, with high practical value, covering code completion [59, 85, 388], code reasoning [90, 325], and code generation [13, 36] tasks. Among these, code generation gains more attention. HumanEval [36] and MBPP [13] provide natural language descriptions of programming problems, requiring LLMs to generate corresponding Python code and use unit tests for evaluation. MultiPL-E [27] extend them to 18 program languages. EvalPlus [187] automatically augments test cases to assess the robustness of the generated code. Recently, some studies collect benchmarks from open-source projects, which are closer to realistic applications and more challenging due to complex function calls, such as DS-1000 [145], CoderEval [369], EvoCodeBench [157] and BigCodeBench [424].

Commonsense reasoning requires LLMs to possess both commonsense and reasoning abilities. StrategyQA [83] collects complex and subtle multi-hop reasoning questions. MMLU [106] and MMLU-Pro [322] cover commonsense reasoning questions across various domains, including STEM, the humanities, the social sciences, etc. *Planning* aims to enable LLMs to take optimal actions based on the current state and environment to complete tasks. Current planning benchmarks primarily focus on synthetic tasks, such as Blocksworld [298], Crosswords, and Game-of-24 [359].

8 Conclusion

In this paper, we conduct a comprehensive survey of existing works on test-time compute. We introduce various test-time compute methods in System-1 and System-2 models, and look forward to future directions for this field. We believe test-time compute can help models handle complex real-world distributions and tasks better, making it a promising path for advancing LLMs toward cognitive intelligence. We hope this paper will promote further research.

References

- [1] Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. 2023. Let’s Sample Step by Step: Adaptive-Consistency for Efficient Reasoning and Coding with LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12375–12396. doi:10.18653/v1/2023.emnlp-main.761
- [2] Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning. arXiv:2503.04697 [cs.CL] <https://arxiv.org/abs/2503.04697>
- [3] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to Basics: Revisiting REINFORCE-Style Optimization for Learning from Human Feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12248–12267. doi:10.18653/v1/2024.acl-long.662

- [4] Ekin Akyürek, Mehul Damani, Linlu Qiu, Han Guo, Yoon Kim, and Jacob Andreas. 2024. The Surprising Effectiveness of Test-Time Training for Abstract Reasoning. arXiv:2411.07279 [cs.AI] <https://arxiv.org/abs/2411.07279>
- [5] Alfonso Amayuelas, Xianjun Yang, Antonis Antoniades, Wenye Hu, Liangming Pan, and William Wang. 2024. MultiAgent Collaboration Attack: Investigating Adversarial Attacks in Large Language Model Collaborations via Debate. arXiv:2406.14711 [cs.CL] <https://arxiv.org/abs/2406.14711>
- [6] Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. Variational Best-of-N Alignment. arXiv:2407.06057 [cs.CL] <https://arxiv.org/abs/2407.06057>
- [7] Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2023. Learning from mistakes makes llm better reasoner. *arXiv preprint arXiv:2310.20689* (2023).
- [8] Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan D. Chang, and Prithviraj Ammanabrolu. 2024. Critique-out-Loud Reward Models. arXiv:2408.11791 [cs.LG] <https://arxiv.org/abs/2408.11791>
- [9] Andy Arditi, Oscar Obeso, Aaqib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in Language Models Is Mediated by a Single Direction. arXiv:2406.11717 [cs.LG] <https://arxiv.org/abs/2406.11717>
- [10] Daman Arora, Himanshu Singh, and Mausam. 2023. Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 7527–7543. doi:10.18653/v1/2023.emnlp-main.468
- [11] Daman Arora and Andrea Zanette. 2025. Training Language Models to Reason Efficiently. arXiv:2502.04463 [cs.LG] <https://arxiv.org/abs/2502.04463>
- [12] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4623–4637. doi:10.18653/v1/2020.acl-main.421
- [13] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732* (2021).
- [14] Simon A. Aytes, Jinheon Baek, and Sung Ju Hwang. 2025. Sketch-of-Thought: Efficient LLM Reasoning with Adaptive Cognitive-Inspired Sketching. arXiv:2503.05179 [cs.CL] <https://arxiv.org/abs/2503.05179>
- [15] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen Marcus McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. Llemma: An Open Language Model for Mathematics. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=4WnqRR915j>
- [16] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).
- [17] Yarin Bar, Shalev Shaer, and Yaniv Romano. 2024. Protected Test-Time Adaptation via Online Entropy Matching: A Betting Approach. arXiv:2408.07511 [cs.LG] <https://arxiv.org/abs/2408.07511>
- [18] Alexander Bartler, Andre Bühler, Felix Wiewel, Mario Döbler, and Bin Yang. 2022. Mt3: Meta test-time training for self-supervised test-time adaption. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3080–3090.
- [19] Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403* (2024).
- [20] Rishabh Bhardwaj, Yingting Li, Navonil Majumder, Bo Cheng, and Soujanya Poria. 2023. kNN-CM: A Non-parametric Inference-Phase Adaptation of Parametric Text Classifiers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13546–13557. doi:10.18653/v1/2023.findings-emnlp.903
- [21] Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. 2024. Forest-of-Thought: Scaling Test-Time Compute for Enhancing LLM Reasoning. arXiv:2412.09078 [cs.CL] <https://arxiv.org/abs/2412.09078>
- [22] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.
- [23] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling. arXiv:2407.21787 [cs.LG] <https://arxiv.org/abs/2407.21787>
- [24] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [25] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. arXiv:2312.09390 [cs.CL] <https://arxiv.org/abs/2312.09390>
- [26] Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized Steering of Large Language Models: Versatile Steering Vectors Through Bi-directional Preference Optimization. arXiv:2406.00045 [cs.CL] <https://arxiv.org/abs/2406.00045>
- [27] Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q Feldman, et al. 2022. Multipl-e: A scalable and extensible approach to benchmarking neural code generation. *arXiv preprint arXiv:2208.08227* (2022).
- [28] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=FQepisCUWu>

- [29] Anirudh Chari, Aditya Tiwari, Richard Lian, Suraj Reddy, and Brian Zhou. 2025. Pheromone-based Learning of Optimal Reasoning Paths. arXiv:2501.19278 [cs.CL] <https://arxiv.org/abs/2501.19278>
- [30] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. 2022. Contrastive Test-Time Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 295–305.
- [31] Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024. AlphaMath Almost Zero: Process Supervision without Process. arXiv:2405.03553 [cs.CL] <https://arxiv.org/abs/2405.03553>
- [32] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025. SFT or RL? An Early Investigation into Training R1-Like Reasoning Large Vision-Language Models. arXiv:2504.11468 [cs.CL] <https://arxiv.org/abs/2504.11468>
- [33] Jiefeng Chen, Jie Ren, Xinyun Chen, Chengrun Yang, Ruoxi Sun, and Sercan Ö Arık. 2025. SETS: Leveraging Self-Verification and Self-Correction for Improved Test-Time Scaling. arXiv:2501.19306 [cs.AI] <https://arxiv.org/abs/2501.19306>
- [34] Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7066–7085. doi:10.18653/v1/2024.acl-long.381
- [35] Lichang Chen, Chen Zhu, Juhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. ODIN: Disentangled Reward Mitigates Hacking in RLHF. In *Forty-first International Conference on Machine Learning*. <https://openreview.net/forum?id=zcIV8OQFVF>
- [36] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
- [37] Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. 2025. JudgeLRM: Large Reasoning Models as a Judge. arXiv:2504.00050 [cs.CL] <https://arxiv.org/abs/2504.00050>
- [38] Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. 2023. Self-ICL: Zero-Shot In-Context Learning with Self-Generated Demonstrations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 15651–15662. doi:10.18653/v1/2023.emnlp-main.968
- [39] Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. 2025. RM-R1: Reward Modeling as Reasoning. arXiv:2505.02387 [cs.CL] <https://arxiv.org/abs/2505.02387>
- [40] Xinyun Chen, Maxwell Lin, Nathanael Schärlí, and Denny Zhou. 2024. Teaching Large Language Models to Self-Debug. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=KuPixlqPiq>
- [41] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuwei Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs. arXiv:2412.21187 [cs.CL] <https://arxiv.org/abs/2412.21187>
- [42] Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. 2025. Reasoning Beyond Language: A Comprehensive Survey on Latent Chain-of-Thought Reasoning. arXiv:2505.16782 [cs.CL] <https://arxiv.org/abs/2505.16782>
- [43] Yanxi Chen, Xuchen Pan, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024. A Simple and Provable Scaling Law for the Test-Time Compute of Large Language Models. arXiv:2411.19477 [cs.CL] <https://arxiv.org/abs/2411.19477>
- [44] Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, Canyu Chen, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhukai Zhao, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. MJ-Bench: Is Your Multimodal Reward Model Really a Good Judge for Text-to-Image Generation? arXiv:2407.04842 [cs.CV] <https://arxiv.org/abs/2407.04842>
- [45] Jeffrey Cheng and Benjamin Van Durme. 2024. Compressed Chain of Thought: Efficient Reasoning Through Dense Representations. arXiv:2412.13171 [cs.CL] <https://arxiv.org/abs/2412.13171>
- [46] Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2025. Think More, Hallucinate Less: Mitigating Hallucinations via Dual Process of Fast and Slow Thinking. arXiv:2501.01306 [cs.CL] <https://arxiv.org/abs/2501.01306>
- [47] Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 15607–15631. doi:10.18653/v1/2023.acl-long.870
- [48] Cheng-Han Chiang, Hung yi Lee, and Michal Lukasik. 2025. TRACT: Regression-Aware Fine-tuning Meets Chain-of-Thought Reasoning for LLM-as-a-Judge. arXiv:2503.04381 [cs.CL] <https://arxiv.org/abs/2503.04381>
- [49] Yinlam Chow, Guy Tennenholtz, Izzeddin Gur, Vincent Zhuang, Bo Dai, Sridhar Thiagarajan, Craig Boutilier, Rishabh Agarwal, Aviral Kumar, and Aleksandra Faust. 2024. Inference-Aware Fine-Tuning for Best-of-N Sampling in Large Language Models. arXiv:2412.15287 [cs.CL] <https://arxiv.org/abs/2412.15287>
- [50] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168 [cs.LG] <https://arxiv.org/abs/2110.14168>
- [51] Wenyan Cong, Hanqing Zhu, Peihao Wang, Bangya Liu, Dejia Xu, Kevin Wang, David Z. Pan, Yan Wang, Zhiwen Fan, and Zhangyang Wang. 2025. Can Test-Time Scaling Improve World Foundation Model? arXiv:2503.24320 [cs.CV] <https://arxiv.org/abs/2503.24320>
- [52] Yingqian Cui, Pengfei He, Jingying Zeng, Hui Liu, Xianfeng Tang, Zhenwei Dai, Yan Han, Chen Luo, Jing Huang, Zhen Li, Suhang Wang, Yue Xing, Jiliang Tang, and Qi He. 2025. Stepwise Perplexity-Guided Refinement for Efficient Chain-of-Thought Reasoning in Large Language Models.

- arXiv:2502.13260 [cs.CL] <https://arxiv.org/abs/2502.13260>
- [53] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta-Optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4005–4019. doi:10.18653/v1/2023.findings-acl.247
- [54] Muzhi Dai, Chenxu Yang, and Qingyi Si. 2025. S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models. arXiv:2505.07686 [cs.AI] <https://arxiv.org/abs/2505.07686>
- [55] Mehul Damani, Idan Shenfeld, Andi Peng, Andreea Bobu, and Jacob Andreas. 2024. Learning How Hard to Think: Input-Adaptive Allocation of LM Computation. arXiv:2410.04707 [cs.LG] <https://arxiv.org/abs/2410.04707>
- [56] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=H1edEyBKDS>
- [57] Google DeepMind. 2025. Gemini 2.5: Our most intelligent AI model. <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/>
- [58] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. OpenVLThinker: An Early Exploration to Complex Vision-Language Reasoning via Iterative Self-Improvement. arXiv:2503.17352 [cs.CV] <https://arxiv.org/abs/2503.17352>
- [59] Yangruibo Ding, Zijian Wang, Wasi Uddin Ahmad, Hantian Ding, Ming Tan, Nihal Jain, Murali Krishna Ramanathan, Ramesh Nallapati, Parminder Bhatia, Dan Roth, and Bing Xiang. 2023. CrossCodeEval: A Diverse and Multilingual Benchmark for Cross-File Code Completion. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. <https://openreview.net/forum?id=wgDcbBMSfh>
- [60] Guanting Dong, Chenghai Zhang, Mengjie Deng, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. 2024. Progressive Multimodal Reasoning via Active Retrieval. arXiv:2412.14835 [cs.CL] <https://arxiv.org/abs/2412.14835>
- [61] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. In *EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1107–1128. doi:10.18653/v1/2024.emnlp-main.64
- [62] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding. *Commun. ACM* 67, 1 (2023), 110–120.
- [63] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325 [cs.CL] <https://arxiv.org/abs/2305.14325>
- [64] Ahmed Elgohary, Christopher Meek, Matthew Richardson, Adam Fourney, Gonzalo Ramos, and Ahmed Hassan Awadallah. 2021. NL-EDIT: Correcting Semantic Parse Errors through Natural Language Interaction. In *NAACL 2021*. Association for Computational Linguistics, Online, 5599–5610. doi:10.18653/v1/2021.nacl-main.444
- [65] Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025. Thinkless: LLM Learns When to Think. arXiv:2505.13379 [cs.CL] <https://arxiv.org/abs/2505.13379>
- [66] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. 2023. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2704–2714.
- [67] Guahao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2024. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems* 36 (2024).
- [68] Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Jingyi Song, and Hao Wang. 2025. AirRAG: Activating Intrinsic Reasoning for Retrieval Augmented Generation using Tree-based Search. arXiv:2501.10053 [cs.AI] <https://arxiv.org/abs/2501.10053>
- [69] Xidong Feng, Ziyu Wan, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. 2024. Alphazero-like Tree-Search can Guide Large Language Model Decoding and Training. arXiv:2309.17179 [cs.LG] <https://arxiv.org/abs/2309.17179>
- [70] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen (Eds.). Association for Computational Linguistics, Hong Kong, China, 1–13. doi:10.18653/v1/D19-5801
- [71] Qichen Fu, Minsik Cho, Thomas Merth, Sachin Mehta, Mohammad Rastegari, and Mahyar Najibi. 2024. Lazylim: Dynamic token pruning for efficient long context llm inference. *arXiv preprint arXiv:2407.14057* (2024).
- [72] Yichao Fu, Junda Chen, Yonghao Zhuang, Zheyu Fu, Ion Stoica, and Hao Zhang. 2025. Reasoning Without Self-Doubt: More Efficient Chain-of-Thought Through Certainty Probing. In *ICLR 2025 Workshop on Foundation Models in the Wild*. <https://openreview.net/forum?id=wpK4IMJfdX>
- [73] Bingzheng Gan, Yufan Zhao, Tianyi Zhang, Jing Huang, Yusu Li, Shu Xian Teo, Changwang Zhang, and Wei Shi. 2025. MASTER: A Multi-Agent System with LLM Specialized MCTS. arXiv:2501.14304 [cs.AI] <https://arxiv.org/abs/2501.14304>
- [74] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A Efros. 2022. Test-Time Training with Masked Autoencoders. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). <https://openreview.net/forum?id=SHMi1b7sjXk>
- [75] Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. 2024. Omni-MATH: A Universal Olympiad Level Mathematic Benchmark For Large Language Models. arXiv:2410.07985 [cs.CL] <https://arxiv.org/abs/2410.07985>
- [76] Ge Gao, Eunsol Choi, and Yoav Artzi. 2022. Simulating Bandit Learning from User Feedback for Extractive Question Answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline

- Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 5167–5179. doi:[10.18653/v1/2022.acl-long.355](https://doi.org/10.18653/v1/2022.acl-long.355)
- [77] Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. 2024. Interleaved-Modal Chain-of-Thought. arXiv:[2411.19488](https://arxiv.org/abs/2411.19488) [cs.CV] <https://arxiv.org/abs/2411.19488>
- [78] Kuofeng Gao, Huanqia Cai, Qingyao Shuai, Dihong Gong, and Zhifeng Li. 2024. Embedding Self-Correction as an Inherent Ability in Large Language Models for Enhanced Mathematical Reasoning. arXiv:[2410.10735](https://arxiv.org/abs/2410.10735) [cs.AI] <https://arxiv.org/abs/2410.10735>
- [79] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and Revising What Language Models Say, Using Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 16477–16508. doi:[10.18653/v1/2023.acl-long.910](https://doi.org/10.18653/v1/2023.acl-long.910)
- [80] Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*. PMLR, 10835–10866.
- [81] Zitian Gao, Boya Niu, Xuzheng He, Haotian Xu, Hongzhang Liu, Aiwei Liu, Xuming Hu, and Lijie Wen. 2024. Interpretable Contrastive Monte Carlo Tree Search Reasoning. arXiv:[2410.01707](https://arxiv.org/abs/2410.01707) [cs.CL] <https://arxiv.org/abs/2410.01707>
- [82] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.
- [83] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics* 9 (2021), 346–361. doi:[10.1162/tacl_a_00370](https://doi.org/10.1162/tacl_a_00370)
- [84] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. arXiv preprint arXiv:[2411.04872](https://arxiv.org/abs/2411.04872) (2024).
- [85] Linyuan Gong, Sida Wang, Mostafa Elhoushi, and Alvin Cheung. 2024. Evaluation of LLMs on Syntax-Aware Code Fill-in-the-Middle Tasks. In *Forty-first International Conference on Machine Learning*. <https://openreview.net/forum?id=jKYyFbH8ap>
- [86] Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. 2024. SoTTA: Robust Test-Time Adaptation on Noisy Data Streams. *Advances in Neural Information Processing Systems* 36 (2024).
- [87] Google. 2024. Try Deep Research and our new experimental model in Gemini, your AI assistant. *Google, blog* (2024). <https://blog.google/products/gemini/google-gemini-deep-research/>
- [88] Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=Sx038qxjek>
- [89] Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnav Nagarajan. 2024. Think before you speak: Training Language Models With Pause Tokens. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=ph04CRkPdC>
- [90] Alex Gu, Baptiste Roziere, Hugh James Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida Wang. 2024. CRUXEval: A Benchmark for Code Reasoning, Understanding and Execution. In *Forty-first International Conference on Machine Learning*. <https://openreview.net/forum?id=Ffpg52swvg>
- [91] Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18126–18134.
- [92] Xinyu Guan, Li Lyra Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. arXiv:[2501.04519](https://arxiv.org/abs/2501.04519) [cs.CL] <https://arxiv.org/abs/2501.04519>
- [93] Lin Gui, Cristina Garbacea, and Victor Veitch. 2024. BoNBoN Alignment for Large Language Models and the Sweetness of Best-of-n Sampling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=haSKMlrBx5>
- [94] Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced Self-Training (ReST) for Language Modeling. arXiv:[2308.08998](https://arxiv.org/abs/2308.08998) [cs.CL] <https://arxiv.org/abs/2308.08998>
- [95] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:[2501.12948](https://arxiv.org/abs/2501.12948) (2025).
- [96] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. 2025. Can We Generate Images with CoT? Let's Verify and Reinforce Image Generation Step by Step. arXiv:[2501.13926](https://arxiv.org/abs/2501.13926) [cs.CV] <https://arxiv.org/abs/2501.13926>
- [97] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science* 3, 10 (2023), 833–838.
- [98] Tingxu Han, Chunrong Fang, Shiyu Zhao, Shiqing Ma, Zhenyu Chen, and Zhenting Wang. 2024. Token-Budget-Aware LLM Reasoning. arXiv:[2412.18547](https://arxiv.org/abs/2412.18547) [cs.CL] <https://arxiv.org/abs/2412.18547>
- [99] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with Language Model is Planning with World Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8154–8173. doi:[10.18653/v1/2023.emnlp-main.507](https://doi.org/10.18653/v1/2023.emnlp-main.507)

- [100] Shibo Hao, Sainbayar Sukhbaatar, Dijia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training Large Language Models to Reason in a Continuous Latent Space. arXiv:2412.06769 [cs.CL] <https://arxiv.org/abs/2412.06769>
- [101] Moritz Hardt and Yu Sun. 2024. Test-Time Training on Nearest Neighbors for Large Language Models. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=CNL2bku4ra>
- [102] Jameel Hassan, Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. 2023. Align your prompts: test-time prompting with distribution alignment for zero-shot generalization. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 80396–80413.
- [103] Alex Havrilla, Sharath Raparthi, Christoforus Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Raileanu. 2024. GLoRe: When, Where, and How to Improve LLM Reasoning via Global and Local Refinements. arXiv:2402.10963 [cs.CL] <https://arxiv.org/abs/2402.10963>
- [104] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 3828–3850. doi:10.18653/v1/2024.acl-long.211
- [105] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8340–8349.
- [106] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. <https://openreview.net/forum?id=7Bywt2mQsCe>
- [107] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=HJz6tiCqYm>
- [108] Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. V-STaR: Training Verifiers for Self-Taught Reasoners. In *First Conference on Language Modeling*. <https://openreview.net/forum?id=stmqBSW2dV>
- [109] Zhenyu Hou, Pengfan Du, Yilin Niu, Zhengxiao Du, Aohan Zeng, Xiao Liu, Minlie Huang, Hongning Wang, Jie Tang, and Yuxiao Dong. 2024. Does RLHF Scale? Exploring the Impacts From Data, Model, and Method. arXiv:2412.06000 [cs.CL] <https://arxiv.org/abs/2412.06000>
- [110] Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. 2025. Advancing Language Model Reasoning through Reinforcement Learning and Inference Scaling. arXiv:2501.11651 [cs.LG] <https://arxiv.org/abs/2501.11651>
- [111] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*. PMLR, 2790–2799.
- [112] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=nZeVKeeFYf9>
- [113] Jinwu Hu, Zhitian Zhang, Guohao Chen, Xutao Wen, Chao Shuai, Wei Luo, Bin Xiao, Yuanqing Li, and Mingkui Tan. 2025. Test-Time Learning for Large Language Models. arXiv:2505.20633 [cs.CL] <https://arxiv.org/abs/2505.20633>
- [114] Minda Hu, Licheng Zong, Hongru Wang, Jingyan Zhou, Jingjing Li, Yichen Gao, Kam-Fai Wong, Yu Li, and Irwin King. 2024. SeRTS: Self-Rewarding Tree Search for Biomedical Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1321–1335. doi:10.18653/v1/2024.findings-emnlp.71
- [115] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. 2024. Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=GNSMI1P5VR>
- [116] Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. 2025. Efficient Test-Time Scaling via Self-Calibration. arXiv:2503.00031 [cs.LG] <https://arxiv.org/abs/2503.00031>
- [117] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large Language Models Cannot Self-Correct Reasoning Yet. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=IkM03fkBPQ>
- [118] Wenzuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models. arXiv:2503.06749 [cs.CV] <https://arxiv.org/abs/2503.06749>
- [119] Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaotong Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. 2024. OlympicArena: Benchmarking Multi-discipline Cognitive Reasoning for Superintelligent AI. arXiv:2406.12753 [cs.CL] <https://arxiv.org/abs/2406.12753>
- [120] Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. 2024. O1 Replication Journey—Part 2: Surpassing O1-preview through Simple Distillation, Big Progress or Bitter Lesson? *arXiv preprint arXiv:2411.16489* (2024).
- [121] Jonas Hübotter, Sascha Bongni, Ido Hakimi, and Andreas Krause. 2025. Efficiently Learning at Test-Time: Active Fine-Tuning of LLMs. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=NS1G1Uhny3>

- [122] Raza Imam, Hanan Gani, Muhammad Huzaifa, and Karthik Nandakumar. 2024. Test-Time Low Rank Adaptation via Confidence Maximization for Zero-Shot Generalization of Vision-Language Models. arXiv:2407.15913 [cs.CV] <https://arxiv.org/abs/2407.15913>
- [123] Chen Jia. 2024. Generalizing Reward Modeling for Out-of-Distribution Preference Learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 107–124.
- [124] Jinbiao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Wayne Xin Zhao, Yang Song, and Tao Zhang. 2024. RAG-Star: Enhancing Deliberative Reasoning with Retrieval Augmented Verification and Refinement. arXiv:2412.12881 [cs.CL] <https://arxiv.org/abs/2412.12881>
- [125] Lingjie Jiang, Xun Wu, Shaohan Huang, Qingxiu Dong, Zewen Chi, Li Dong, Xingxing Zhang, Tengchao Lv, Lei Cui, and Furu Wei. 2025. Think Only When You Need with Large Hybrid-Reasoning Models. arXiv:2505.14631 [cs.CL] <https://arxiv.org/abs/2505.14631>
- [126] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. arXiv:2503.09516 [cs.CL] <https://arxiv.org/abs/2503.09516>
- [127] Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux* (2011).
- [128] Ryō Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. arXiv:2406.01297 [cs.CL] <https://arxiv.org/abs/2406.01297>
- [129] Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. 2025. C3ot: Generating shorter chain-of-thought without compromising effectiveness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 24312–24320.
- [130] Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable Best-of-N Selection for Large Language Models via Self-Certainty. arXiv:2502.18581 [cs.CL] <https://arxiv.org/abs/2502.18581>
- [131] Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. 2025. Process Reward Models That Think. arXiv:2504.16828 [cs.LG] <https://arxiv.org/abs/2504.16828>
- [132] Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with More Persuasive LLMs Leads to More Truthful Answers. In *Forty-first International Conference on Machine Learning*. <https://openreview.net/forum?id=iLCZtl7FTa>
- [133] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest Neighbor Machine Translation. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=7wCBOfj8hJM>
- [134] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=HkdlBjCEKvH>
- [135] Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language Models can Solve Computer Tasks. arXiv:2303.17491 [cs.CL] <https://arxiv.org/abs/2303.17491>
- [136] Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang goo Lee. 2022. Self-Generated In-Context Learning: Leveraging Auto-regressive Language Models as a Demonstration Generator. arXiv:2206.08082 [cs.CL] <https://arxiv.org/abs/2206.08082>
- [137] Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can LLMs Produce Faithful Explanations For Fact-checking? Towards Faithful Explainable Fact-Checking via Multi-Agent Debate. arXiv:2402.07401 [cs.CL] <https://arxiv.org/abs/2402.07401>
- [138] Sungewan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Jungsoo Won, Dongha Lee, and Jinyoung Yeo. 2024. Evaluating Robustness of Reward Models for Mathematical Reasoning. *arXiv preprint arXiv:2410.01729* (2024).
- [139] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=8euJaTveKw>
- [140] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 4334–4353. doi:10.18653/v1/2024.emnlp-main.248
- [141] Seungone Kim, Ian Wu, Jinu Lee, Xiang Yue, Seongyun Lee, Mingyeong Moon, Kiril Gashtevski, Carolin Lawrence, Julia Hockenmaier, Graham Neubig, and Sean Welleck. 2025. Scaling Evaluation-time Compute with Reasoning Models as Process Evaluators. arXiv:2503.19877 [cs.CL] <https://arxiv.org/abs/2503.19877>
- [142] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.
- [143] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. 2024. Training Language Models to Self-Correct via Reinforcement Learning. arXiv:2409.12917 [cs.LG] <https://arxiv.org/abs/2409.12917>
- [144] Philippe Laban, Lidiya Murakhov's'ka, Caiming Xiong, and Chien-Sheng Wu. 2024. Are You Sure? Challenging LLMs Leads to Performance Drops in The FlipFlop Experiment. arXiv:2311.08596 [cs.CL] <https://arxiv.org/abs/2311.08596>
- [145] Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-tau Yih, Daniel Fried, Sida Wang, and Tao Yu. 2023. DS-1000: A natural and reliable benchmark for data science code generation. In *International Conference on Machine Learning*. PMLR, 18319–18345.
- [146] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. RewardBench: Evaluating Reward Models for Language Modeling. arXiv:2403.13787 [cs.LG] <https://arxiv.org/abs/2403.13787>

- [147] Tian Lan, Wenwei Zhang, Chengqi Lyu, Shuaibin Li, Chen Xu, Heyan Huang, Dahua Lin, Xian-Ling Mao, and Kai Chen. 2024. Training language models to critique with multi-agent feedback. *arXiv preprint arXiv:2410.15287* (2024).
- [148] Yann LeCun. 2022. A path towards autonomous machine intelligence. (2022).
- [149] Ayeong Lee, Ethan Che, and Tianyi Peng. 2025. How Well do LLMs Compress Their Own Chain-of-Thought? A Token Complexity Approach. *arXiv:2503.01141* [cs.CL] <https://arxiv.org/abs/2503.01141>
- [150] Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. 2024. Multimodal Reasoning with Multimodal Knowledge Graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 10767–10782. doi:10.18653/v1/2024.acl-long.579
- [151] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 3045–3059. doi:10.18653/v1/2021.emnlp-main.243
- [152] Noam Levi. 2024. A Simple Model of Inference Scaling Laws. *arXiv preprint arXiv:2410.16377* (2024).
- [153] Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*. PMLR, 19274–19286.
- [154] Will LeVine, Benjamin Plikus, Anthony Chen, and Sean Hendryx. 2023. A Baseline Analysis of Reward Models' Ability To Accurately Analyze Foundation Models Under Distribution Shift. *arXiv preprint arXiv:2311.14743* (2023).
- [155] Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7315–7330. doi:10.18653/v1/2020.acl-main.653
- [156] Jiachun Li, Pengfei Cao, Yubo Chen, Jiebin Xu, Huajun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2025. Rewarding Curse: Analyze and Mitigate Reward Modeling Issues for LLM Reasoning. *arXiv:2503.05188* [cs.CL] <https://arxiv.org/abs/2503.05188>
- [157] Jia Li, Ge Li, Xuanming Zhang, Yihong Dong, and Zhi Jin. 2024. EvoCodeBench: An Evolving Code Generation Benchmark Aligned with Real-World Code Repositories. *arXiv preprint arXiv:2404.00599* (2024).
- [158] Junyi Li and Hwee Tou Ng. 2024. Think&Cite: Improving Attributed Text Generation with Self-Guided Tree Search and Progress Reward Modeling. *arXiv:2412.14860* [cs.CL] <https://arxiv.org/abs/2412.14860>
- [159] Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, hai zhao, and Pengfei Liu. 2024. Generative Judge for Evaluating Alignment. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=gtkFw6sZGS>
- [160] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. More Agents Is All You Need. *arXiv:2402.05120* [cs.CL] <https://arxiv.org/abs/2402.05120>
- [161] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=aLLuYpn83y>
- [162] Loka Li, Zhenhao Chen, Guangyi Chen, Yixuan Zhang, Yusheng Su, Eric Xing, and Kun Zhang. 2024. Confidence Matters: Revisiting Intrinsic Self-Correction Capabilities of Large Language Models. *arXiv:2402.12563* [cs.CL] <https://arxiv.org/abs/2402.12563>
- [163] Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. 2024. VLRewardBench: A Challenging Benchmark for Vision-Language Generative Reward Models. *arXiv:2411.17451* [cs.CV] <https://arxiv.org/abs/2411.17451>
- [164] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic Search-Enhanced Large Reasoning Models. *arXiv:2501.05366* [cs.AI] <https://arxiv.org/abs/2501.05366>
- [165] Xiang Li, Shizhu He, Jiayu Wu, Zhao Yang, Yao Xu, Yang jun Jun, Haifeng Liu, Kang Liu, and Jun Zhao. 2024. MoDE-CoTD: Chain-of-Thought Distillation for Complex Reasoning Tasks with Mixture of Decoupled LoRA-Experts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 11475–11485. <https://aclanthology.org/2024.lrec-main.1003/>
- [166] Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025. WebThinker: Empowering Large Reasoning Models with Deep Research Capability. *arXiv:2504.21776* [cs.CL] <https://arxiv.org/abs/2504.21776>
- [167] Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified Demonstration Retriever for In-Context Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4644–4668. doi:10.18653/v1/2023.acl-long.256
- [168] Xiaonan Li and Xipeng Qiu. 2023. Finding Support Examples for In-Context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6219–6235. doi:10.18653/v1/2023.findings-emnlp.411
- [169] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittweis, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science* 378, 6624 (2022), 1092–1097.
- [170] Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving Multi-Agent Debate with Sparse Communication Topology. *arXiv:2406.11776* [cs.CL] <https://arxiv.org/abs/2406.11776>

- [171] Yafu Li, Xuyang Hu, Xiaoye Qu, Linjie Li, and Yu Cheng. 2025. Test-Time Preference Optimization: On-the-Fly Alignment via Iterative Textual Feedback. arXiv:2501.12895 [cs.CL] <https://arxiv.org/abs/2501.12895>
- [172] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making Language Models Better Reasoners with Step-Aware Verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5315–5333. doi:10.18653/v1/2023.acl-long.291
- [173] Zheng Li, Qingxiu Dong, Jingyuan Ma, Di Zhang, and Zhifang Sui. 2025. SelfBudgeter: Adaptive Token Allocation for Efficient LLM Reasoning. arXiv:2505.11274 [cs.AI] <https://arxiv.org/abs/2505.11274>
- [174] Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. 2024. Chain of Thought Empowers Transformers to Solve Inherently Serial Problems. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=3EWTEy9MTM>
- [175] Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie Cheung, and Siva Reddy. 2022. Using Interactive Feedback to Improve the Accuracy and Explainability of Question Answering Systems Post-Deployment. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Alina Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 926–937. doi:10.18653/v1/2022.findings-acl.75
- [176] Zhen Li, Yupeng Su, Running Yang, Zhongwei Xie, Ngai Wong, and Hongxia Yang. 2025. Quantization Meets Reasoning: Exploring LLM Low-Bit Quantization Degradation for Mathematical Reasoning. arXiv preprint arXiv:2501.03035 (2025).
- [177] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. arXiv:2305.19118 [cs.CL] <https://arxiv.org/abs/2305.19118>
- [178] Xiaobo Liang, Haoke Zhang, Helan hu, Juntao Li, Jun Xu, and Min Zhang. 2024. Fennec: Fine-grained Language Model Evaluation and Correction Extended through Branching and Bridging. arXiv:2405.12163 [cs.CL] <https://arxiv.org/abs/2405.12163>
- [179] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's Verify Step by Step. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=v8L0pN6EOi>
- [180] Lei Lin, Jiayi Fu, Pengli Liu, Qingyang Li, Yan Gong, Junchen Wan, Fuzheng Zhang, Zhongyuan Wang, Di Zhang, and Kun Gai. 2024. Just Ask One More Time! Self-Agreement Improves Reasoning of Language Models in (Almost) All Scenarios. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 3829–3852. doi:10.18653/v1/2024.findings-acl.230
- [181] Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Yun-Nung Chen. 2024. DogeRM: Equipping Reward Models with Domain Knowledge through Model Merging. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 15506–15524. doi:10.18653/v1/2024.emnlp-main.868
- [182] Yujie Lin, Ante Wang, Moye Chen, Jingyao Liu, Hao Liu, Jinsong Su, and Xinyan Xiao. 2025. Investigating Inference-time Scaling for Chain of Multi-modal Thought: A Preliminary Study. arXiv:2502.11514 [cs.CL] <https://arxiv.org/abs/2502.11514>
- [183] Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024. CriticBench: Benchmarking LLMs for Critique-Correct Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 1552–1587. doi:10.18653/v1/2024.findings-acl.91
- [184] Fan Liu, Wenshuo Chao, Naiqiang Tan, and Hao Liu. 2025. Bag of Tricks for Inference-time Computation of LLM Reasoning. arXiv:2502.07191 [cs.AI] <https://arxiv.org/abs/2502.07191>
- [185] Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. 2025. Video-T1: Test-Time Scaling for Video Generation. arXiv:2503.18942 [cs.CV] <https://arxiv.org/abs/2503.18942>
- [186] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3?. In *Proceedings of Deep Learning Inside Out (DeelIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, Eneko Agirre, Marianna Apidianaki, and Ivan Vulić (Eds.). Association for Computational Linguistics, Dublin, Ireland and Online, 100–114. doi:10.18653/v1/2022.deelio-1.10
- [187] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2024. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems* 36 (2024).
- [188] Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. In-context Vectors: Making In Context Learning More Effective and Controllable Through Latent Space Steering. arXiv:2311.06668 [cs.LG] <https://arxiv.org/abs/2311.06668>
- [189] Tengxiao Liu, Qipeng Guo, Xiangkun Hu, Cheng Jiayang, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2024. Can Language Models Learn to Skip Steps?. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 45359–45385. https://proceedings.neurips.cc/paper_files/paper/2024/file/504fa7e518da9d1b53a233ed20a38b46-Paper-Conference.pdf
- [190] Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. 2024. GroupDebate: Enhancing the Efficiency of Multi-Agent Debate Using Group Discussion. arXiv:2409.14051 [cs.CL] <https://arxiv.org/abs/2409.14051>
- [191] Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2511–2522. doi:10.18653/v1/2023.emnlp-main.153
- [192] Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. 2021. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems* 34 (2021), 21808–21820.

- [193] Yuliang Liu, Junjie Lu, Zhaoling Chen, Chaofeng Qu, Jason Klein Liu, Chonghan Liu, Zefan Cai, Yunhui Xia, Li Zhao, Jiang Bian, Chuheng Zhang, Wei Shen, and Zhouhan Lin. 2025. AdaptiveStep: Automatically Dividing Reasoning Step through Model Confidence. arXiv:2502.13943 [cs.AI] <https://arxiv.org/abs/2502.13943>
- [194] Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024. RM-bench: Benchmarking reward models of language models with subtlety and style. *arXiv preprint arXiv:2410.16184* (2024).
- [195] Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2025. Pairwise RM: Perform Best-of-N Sampling with Knockout Tournament. *arXiv preprint arXiv:2501.13007* (2025).
- [196] Zihan Liu, Yang Chen, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. AceMath: Advancing Frontier Math Reasoning with Post-Training and Reward Modeling. arXiv:2412.15084 [cs.CL] <https://arxiv.org/abs/2412.15084>
- [197] Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-Time Scaling for Generalist Reward Modeling. arXiv:2504.02495 [cs.CL] <https://arxiv.org/abs/2504.02495>
- [198] Jieyi Long. 2023. Large Language Model Guided Tree-of-Thought. arXiv:2305.08291 [cs.AI] <https://arxiv.org/abs/2305.08291>
- [199] Chenwei Lou, Zewei Sun, Xinnian Liang, Meng Qu, Wei Shen, Wenqi Wang, Yuntao Li, Qingping Yang, and Shuangzhi Wu. 2025. AdaCoT: Pareto-Optimal Adaptive Chain-of-Thought Triggering via Reinforcement Learning. arXiv:2505.11896 [cs.LG] <https://arxiv.org/abs/2505.11896>
- [200] Dawn Lu and Nina Rimsky. 2024. Investigating Bias Representations in Llama 2 Chat via Activation Steering. arXiv:2402.00402 [cs.CL] <https://arxiv.org/abs/2402.00402>
- [201] Jianqiao Lu, Zhiyang Dou, Hongru WANG, Zeyu Cao, Jianbo Dai, Yunlong Feng, and Zhijiang Guo. 2024. AutoPSV: Automated Process-Supervised Verifier. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=eOAPWOGs9>
- [202] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 8086–8098. doi:10.18653/v1/2022.acl-long.556
- [203] Haoran Luo, Haihong E, Yikai Guo, Qika Lin, Xiaobao Wu, Xinyu Mu, Wenhao Liu, Meina Song, Yifan Zhu, and Luu Anh Tuan. 2025. KBQA-o1: Agentic Knowledge Base Question Answering with Monte Carlo Tree Search. arXiv:2501.18922 [cs.CL] <https://arxiv.org/abs/2501.18922>
- [204] Haotian Luo, Haiying He, Yibo Wang, Jinluan Yang, Rui Liu, Naiqiang Tan, Xiaochun Cao, Dacheng Tao, and Li Shen. 2025. Ada-R1: Hybrid-CoT via Bi-Level Adaptive Reasoning Optimization. arXiv:2504.21659 [cs.AI] <https://arxiv.org/abs/2504.21659>
- [205] Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025. O1-Pruner: Length-Harmonizing Fine-Tuning for O1-Like Reasoning Pruning. arXiv:2501.12570 [cs.CL] <https://arxiv.org/abs/2501.12570>
- [206] Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. 2024. Improve Mathematical Reasoning in Language Models by Automated Process Supervision. arXiv:2406.06592 [cs.CL] <https://arxiv.org/abs/2406.06592>
- [207] Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao. 2023. Dr.ICL: Demonstration-Retrieved In-context Learning. arXiv:2305.14128 [cs.CL] <https://arxiv.org/abs/2305.14128>
- [208] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. ChatGPT as a Factual Inconsistency Evaluator for Text Summarization. arXiv:2303.15621 [cs.CL] <https://arxiv.org/abs/2303.15621>
- [209] Xinxin Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Z-ICL: Zero-Shot In-Context Learning with Pseudo-Demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 2304–2317. doi:10.18653/v1/2023.acl-long.129
- [210] Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. Reasoning Models Can Be Effective Without Thinking. arXiv:2504.09858 [cs.AI] <https://arxiv.org/abs/2504.09858>
- [211] Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2025. CoT-Valve: Length-Compressible Chain-of-Thought Tuning. arXiv:2502.09601 [cs.AI] <https://arxiv.org/abs/2502.09601>
- [212] XIAOSONG MA, Jie ZHANG, Song Guo, and Wenchao Xu. 2023. SwapPrompt: Test-Time Prompt Adaptation for Vision-Language Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 65252–65264. https://proceedings.neurips.cc/paper_files/paper/2023/file/cdd0640218a27e9e2c0e52e324e25db0-Paper-Conference.pdf
- [213] Yan Ma, Steffi Chern, Xuyang Shen, Yiran Zhong, and Pengfei Liu. 2025. Rethinking RL Scaling for Vision Language Models: A Transparent, From-Scratch Framework and Comprehensive Evaluation Scheme. arXiv:2504.02587 [cs.LG] <https://arxiv.org/abs/2504.02587>
- [214] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=S37hOerQLB>
- [215] Rohin Manvi, Anikait Singh, and Stefano Ermon. 2024. Adaptive Inference-Time Compute: LLMs Can Predict if They Can Do Better, Even Mid-Generation. arXiv:2410.02725 [cs.CL] <https://arxiv.org/abs/2410.02725>

- [216] Silin Meng, Yiwei Wang, Cheng-Fu Yang, Nanyun Peng, and Kai-Wei Chang. 2024. LLM-A*: Large Language Model Enhanced Incremental Heuristic Search on Path Planning. arXiv:2407.02511 [cs.RO] <https://arxiv.org/abs/2407.02511>
- [217] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18798–18806.
- [218] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. arXiv:2501.19393 [cs.CL] <https://arxiv.org/abs/2501.19393>
- [219] Dilxat Muhtar, Yelong Shen, Yaming Yang, Xiaodong Liu, Yadong Lu, Jianfeng Liu, Yuefeng Zhan, Hao Sun, Weiwei Deng, Feng Sun, Xueliang Zhang, Jianfeng Gao, Weizhu Chen, and Qi Zhang. 2024. StreamAdapter: Efficient Test Time Adaptation from Contextual Streams. arXiv:2411.09289 [cs.CL] <https://arxiv.org/abs/2411.09289>
- [220] Tergel Munkhbat, Namgyu Ho, Seo Hyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. 2025. Self-Training Elicits Concise Reasoning in Large Language Models. arXiv:2502.20122 [cs.CL] <https://arxiv.org/abs/2502.20122>
- [221] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted question-answering with human feedback. arXiv:2112.09332 [cs.CL] <https://arxiv.org/abs/2112.09332>
- [222] Deepak Nathani, David Wang, Liangming Pan, and William Wang. 2023. MAF: Multi-Aspect Feedback for Improving Reasoning in Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6591–6616. doi:10.18653/v1/2023.emnlp-main.407
- [223] Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. 2024. Test-Time Model Adaptation with Only Forward Passes. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 38298–38315. <https://proceedings.mlr.press/v235/niu24a.html>
- [224] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiqian Wen, Yaofu Chen, Peilin Zhao, and Mingkui Tan. 2023. Towards Stable Test-time Adaptation in Dynamic Wild World. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=g2YraF75Tj>
- [225] Theo X. Olaussen, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. 2024. Is Self-Repair a Silver Bullet for Code Generation?. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=y0GJXRungR>
- [226] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023).
- [227] OpenAI. 2024. Learning to Reason with LLMs. *Open AI, blog* (2024).
- [228] OpenAI. 2025. Introducing deep research. *Open AI, blog* (2025). <https://openai.com/index/introducing-deep-research/>
- [229] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] <https://arxiv.org/abs/2203.02155>
- [230] Sungjin Park, Xiao Liu, Yeyun Gong, and Edward Choi. 2024. Ensembling Large Language Models with Process Reward-Guided Tree Search for Better Complex Reasoning. *arXiv preprint arXiv:2412.15797* (2024).
- [231] Debit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2024. REFINER: Reasoning Feedback on Intermediate Representations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 1100–1126. <https://aclanthology.org/2024.eacl-long.67>
- [232] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.
- [233] Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Bin Xu, Lei Hou, and Juanzi Li. 2025. Agentic Reward Modeling: Integrating Human Preferences with Verifiable Correctness Signals for Reliable Reward Systems. arXiv:2502.19328 [cs.CL] <https://arxiv.org/abs/2502.19328>
- [234] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. LMM-R1: Empowering 3B LMMs with Strong Reasoning Abilities Through Two-Stage Rule-Based RL. arXiv:2503.07536 [cs.CL] <https://arxiv.org/abs/2503.07536>
- [235] Ori Press, Ravid Schwartz-Ziv, Yann LeCun, and Matthias Bethge. 2024. The Entropy Enigma: Success and Failure of Entropy Minimization. arXiv:2405.05012 [cs.CV] <https://arxiv.org/abs/2405.05012>
- [236] Jianing Qi, Hao Tang, and Zhigang Zhu. 2024. VerifierQ: Enhancing LLM Test Time Compute with Q-Learning-based Verifiers. arXiv:2410.08048 [cs.LG] <https://arxiv.org/abs/2410.08048>
- [237] Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenzuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. 2025. VCR-Bench: A Comprehensive Evaluation Framework for Video Chain-of-Thought Reasoning. arXiv:2504.07956 [cs.CV] <https://arxiv.org/abs/2504.07956>
- [238] Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyra Zhang, Fan Yang, and Mao Yang. 2024. Mutual Reasoning Makes Smaller LLMs Stronger Problem-Solvers. arXiv:2408.06195 [cs.CL] <https://arxiv.org/abs/2408.06195>
- [239] Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. 2024. In-Context Learning with Iterative Demonstration Selection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 7441–7455. doi:10.18653/v1/2024.findings-emnlp.438

- [240] Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. 2024. O1 Replication Journey: A Strategic Progress Report – Part 1. arXiv:2410.18982 [cs.AI] <https://arxiv.org/abs/2410.18982>
- [241] Jiahao Qiu, Yifu Lu, Yifan Zeng, Jiacheng Guo, Jiayi Geng, Huazheng Wang, Kaixuan Huang, Yue Wu, and Mengdi Wang. 2024. TreeBoN: Enhancing Inference-Time Alignment with Speculative Tree-Search and Best-of-N Sampling. arXiv:2410.16033 [cs.CL] <https://arxiv.org/abs/2410.16033>
- [242] Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay B Cohen. 2024. Spectral Editing of Activations for Large Language Model Alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=pqYceEa87j>
- [243] Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. 2024. Recursive Introspection: Teaching Language Model Agents How to Self-Improve. arXiv:2407.18219 [cs.LG] <https://arxiv.org/abs/2407.18219>
- [244] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html>
- [245] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=HPuSIXJaa9>
- [246] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2685–2702. doi:10.18653/v1/2020.emnlp-main.213
- [247] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15504–15522. doi:10.18653/v1/2024.acl-long.828
- [248] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémie Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. Code Llama: Open Foundation Models for Code. arXiv:2308.12950 [cs.CL] <https://arxiv.org/abs/2308.12950>
- [249] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning To Retrieve Prompts for In-Context Learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 2655–2671. doi:10.18653/v1/2022.naacl-main.191
- [250] Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-Solve-Merge Improves Large Language Model Evaluation and Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 8352–8370. doi:10.18653/v1/2024.naacl-long.462
- [251] Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. 2025. Learning to Plan & Reason for Evaluation with Thinking-LLM-as-a-Judge. arXiv:2501.18099 [cs.AI] <https://arxiv.org/abs/2501.18099>
- [252] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. arXiv:2206.05802 [cs.CL] <https://arxiv.org/abs/2206.05802>
- [253] Daniel Scalena, Gabriele Sarti, and Malvina Nissim. 2024. Multi-property Steering of Large Language Models with Dynamic Activation Composition. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (Eds.). Association for Computational Linguistics, Miami, Florida, US, 577–603. doi:10.18653/v1/2024.blackboxnlp-1.34
- [254] Alexander Scarlatos and Andrew Lan. 2024. RetICL: Sequential Retrieval of In-Context Examples with Reinforcement Learning. arXiv:2305.14502 [cs.CL] <https://arxiv.org/abs/2305.14502>
- [255] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. 2020. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems* 33 (2020), 11539–11551.
- [256] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 6490–6500. doi:10.18653/v1/2021.acl-long.507
- [257] Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shariari, Sarah Perrin, Abe Friesen, Geoffrey Cideron, Sertan Girgin, Piotr Stanczyk, Andrea Michi, Danila Sinopalnikov, Sabela Ramos, Amélie Héliou, Aliaksei Severyn, Matt Hoffman, Nikola Momchev, and Olivier Bachem. 2024. BOND: Aligning LLMs with Best-of-N Distillation. arXiv:2407.14622 [cs.LG] <https://arxiv.org/abs/2407.14622>
- [258] Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2024. Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning. arXiv:2410.08146 [cs.LG] <https://arxiv.org/abs/2410.08146>

- [259] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs.CL] <https://arxiv.org/abs/2402.03300>
- [260] Shuaijie She, Junxiao Liu, Yifeng Liu, Jiajun Chen, Xin Huang, and Shujian Huang. 2025. R-PRM: Reasoning-Driven Process Reward Modeling. arXiv:2503.21295 [cs.CL] <https://arxiv.org/abs/2503.21295>
- [261] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiansheng Zhao. 2025. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. arXiv:2504.07615 [cs.CV] <https://arxiv.org/abs/2504.07615>
- [262] Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025. CODI: Compressing Chain-of-Thought into Continuous Space via Self-Distillation. arXiv:2502.21074 [cs.CL] <https://arxiv.org/abs/2502.21074>
- [263] Noah Shimm, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366 [cs.AI] <https://arxiv.org/abs/2303.11366>
- [264] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems* 35 (2022), 14274–14289.
- [265] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. arXiv:2408.03314 [cs.LG] <https://arxiv.org/abs/2408.03314>
- [266] Guijin Son, Jiwoo Hong, Hyunwoo Ko, and James Thorpe. 2025. Linguistic Generalizability of Test-Time Scaling in Mathematical Reasoning. arXiv:2502.17407 [cs.CL] <https://arxiv.org/abs/2502.17407>
- [267] Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. PRMBench: A Fine-grained and Challenging Benchmark for Process-Level Reward Models. *arXiv preprint arXiv:2501.03124* (2025).
- [268] Xiaoshuai Song, Yanan Wu, Weixun Wang, Jiaheng Liu, Wenbo Su, and Bo Zheng. 2025. ProgCo: Program Helps Self-Correction of Large Language Models. arXiv:2501.01264 [cs.CL] <https://arxiv.org/abs/2501.01264>
- [269] Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. arXiv:2409.12183 [cs.CL] <https://arxiv.org/abs/2409.12183>
- [270] Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. GPT-4 Doesn't Know It's Wrong: An Analysis of Iterative Prompting for Reasoning Problems. arXiv:2310.12397 [cs.AI] <https://arxiv.org/abs/2310.12397>
- [271] Kaya Stechly, Karthik Valmecikam, and Subbarao Kambhampati. 2024. Chain of Thoughtlessness? An Analysis of CoT in Planning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=kPBEAZU5Nm>
- [272] Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. 2024. Teaching Models to Balance Resisting and Accepting Persuasion. arXiv:2410.14596 [cs.CL] <https://arxiv.org/abs/2410.14596>
- [273] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 3008–3021. https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf
- [274] Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2024. Improving Instruction-Following in Language Models through Activation Steering. arXiv:2410.12877 [cs.CL] <https://arxiv.org/abs/2410.12877>
- [275] Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large Language Models are Inconsistent and Biased Evaluators. arXiv:2405.01724 [cs.CL] <https://arxiv.org/abs/2405.01724>
- [276] Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2022. Selective Annotation Makes Language Models Better Few-Shot Learners. arXiv:2209.01975 [cs.CL] <https://arxiv.org/abs/2209.01975>
- [277] Yi Su, Yixin Ji, Juntao Li, Hai Ye, and Min Zhang. 2023. Beware of Model Collapse! Fast and Stable Test-time Adaptation for Robust Question Answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12998–13011. doi:[10.18653/v1/2023.emnlp-main.803](https://doi.org/10.18653/v1/2023.emnlp-main.803)
- [278] Yi Su, Yunpeng Tai, Yixin Ji, Juntao Li, Yan Bowen, and Min Zhang. 2024. Demonstration Augmentation for Zero-shot In-context Learning. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 14232–14244. doi:[10.18653/v1/2024.findings-acl.846](https://doi.org/10.18653/v1/2024.findings-acl.846)
- [279] Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025. Crossing the Reward Bridge: Expanding RL with Verifiable Rewards Across Diverse Domains. arXiv:2503.23829 [cs.CL] <https://arxiv.org/abs/2503.23829>
- [280] Vighnesh Subramanian, Yilun Du, Joshua B. Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. 2025. Multiagent Finetuning of Language Models. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=JtGP1ZpOrz>
- [281] Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. 2024. Fast Best-of-N Decoding via Speculative Rejection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=348hfcpnUs>
- [282] Wei Sun, Qianlong Du, Fuwei Cui, and Jiajun Zhang. 2025. An Efficient and Precise Training Data Construction Framework for Process-supervised Reward Model in Mathematical Reasoning. arXiv:2503.02382 [cs.CL] <https://arxiv.org/abs/2503.02382>

- [283] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*. PMLR, 9229–9248.
- [284] Niket Tandon, Aman Madaan, Peter Clark, Keisuke Sakaguchi, and Yiming Yang. 2021. Interscript: A dataset for interactive learning of scripts through error feedback. arXiv:2112.07867 [cs.AI] <https://arxiv.org/abs/2112.07867>
- [285] Niket Tandon, Aman Madaan, Peter Clark, and Yiming Yang. 2022. Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 339–352. doi:10.18653/v1/2022.findings-naacl.26
- [286] Xinyu Tang, Xiaolei Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2024. DAWN-ICL: Strategic Planning of Problem-solving Trajectories for Zero-Shot In-Context Learning. *arXiv preprint arXiv:2410.20215* (2024).
- [287] Zhengyang Tang, Ziniu Li, Zhenyang Xiao, Tian Ding, Ruoyu Sun, Benyou Wang, Dayiheng Liu, Fei Huang, Tianyu Liu, Bowen Yu, and Junyang Lin. 2025. Enabling Scalable Oversight via Self-Evolving Critic. arXiv:2501.05727 [cs.CL] <https://arxiv.org/abs/2501.05727>
- [288] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599* (2025).
- [289] Keyu Tian, Yi Jiang, Zehuan Yuan, BINGYUE PENG, and Liwei Wang. 2024. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=gojL67Cf8>
- [290] Vernon Y. H. Toh, Deepanway Ghosal, and Soujanya Poria. 2024. Not All Votes Count! Programs as Verifiers Improve Self-Consistency of Language Models for Math Reasoning. arXiv:2410.12608 [cs.CL] <https://arxiv.org/abs/2410.12608>
- [291] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature* 625, 7995 (2024), 476–482.
- [292] Harsh Trivedi, Niranjana Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 10014–10037. doi:10.18653/v1/2023.acl-long.557
- [293] Haoqin Tu, Weitao Feng, Hardy Chen, Hui Liu, Xianfeng Tang, and Cihang Xie. 2025. ViLBench: A Suite for Vision-Language Process Reward Modeling. arXiv:2503.20271 [cs.CV] <https://arxiv.org/abs/2503.20271>
- [294] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering Language Models With Activation Engineering. arXiv:2308.10248 [cs.CL] <https://arxiv.org/abs/2308.10248>
- [295] Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. 2024. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 13894–13908. doi:10.18653/v1/2024.findings-acl.826
- [296] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. arXiv:2211.14275 [cs.LG] <https://arxiv.org/abs/2211.14275>
- [297] Karthik Valmecam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can Large Language Models Really Improve by Self-critiquing Their Own Plans? arXiv:2310.08118 [cs.AI] <https://arxiv.org/abs/2310.08118>
- [298] Karthik Valmecam, Matthew Marquez, Alberto Olmo, Sarah Sreedharan, and Subbarao Kambhampati. 2023. PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. <https://openreview.net/forum?id=YXogl4uQUO>
- [299] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbahn. 2024. Will we run out of data? Limits of LLM scaling based on human-generated data. arXiv:2211.04325 [cs.LG] <https://arxiv.org/abs/2211.04325>
- [300] Chaojie Wang, Yanchen Deng, Zhiyi Lyu, Liang Zeng, Jujie He, Shuicheng Yan, and Bo An. 2024. Q*: Improving Multi-step Reasoning for LLMs with Deliberative Planning. arXiv:2406.14283 [cs.AI] <https://arxiv.org/abs/2406.14283>
- [301] Dexin Wang, Kai Fan, Boxing Chen, and Deyi Xiong. 2022. Efficient Cluster-Based k -Nearest-Neighbor Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 2175–2187. doi:10.18653/v1/2022.acl-long.154
- [302] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=uXl3bZLkr3c>
- [303] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems* 32 (2019).
- [304] Hanlin Wang, Chak Tou Leong, Jian Wang, and Wenjie Li. 2024. E²CL: Exploration-based Error Correction Learning for Embodied Agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 7626–7639. doi:10.18653/v1/2024.findings-emnlp.448
- [305] Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. *arXiv preprint arXiv:2406.12845* (2024).
- [306] Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, and Ben Athiwaratkun. 2024. Reasoning in Token Economies: Budget-Aware Evaluation of LLM Reasoning Strategies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Manuscript submitted to ACM

- Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 19916–19939. doi:[10.18653/v1/2024.emnlp-main.1112](https://doi.org/10.18653/v1/2024.emnlp-main.1112)
- [307] Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, and Ben Athiwaratkun. 2024. Reasoning in Token Economies: Budget-Aware Evaluation of LLM Reasoning Strategies. arXiv:[2406.06461](https://arxiv.org/abs/2406.06461) [cs.CL] <https://arxiv.org/abs/2406.06461>
- [308] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, Yue Dong, Wen Xiao, Lu Wang, Fei Liu, and Giuseppe Carenini (Eds.). Association for Computational Linguistics, Singapore, 1–11. doi:[10.18653/v1/2023.newsum-1.1](https://doi.org/10.18653/v1/2023.newsum-1.1)
- [309] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-Agents Enhances Large Language Model Capabilities. arXiv:[2406.04692](https://arxiv.org/abs/2406.04692) [cs.CL] <https://arxiv.org/abs/2406.04692>
- [310] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 9440–9450. doi:[10.18653/v1/2024.acl-long.511](https://doi.org/10.18653/v1/2024.acl-long.511)
- [311] Peiyi Wang, Lei Li, Zihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 9426–9439. doi:[10.18653/v1/2024.acl-long.510](https://doi.org/10.18653/v1/2024.acl-long.510)
- [312] Shuhu Wang, Xiaoya Li, Yuxian Meng, Tianwei Zhang, Rongbin Ouyang, Jiwei Li, and Guoyin Wang. 2022. *k*NN-NER: Named Entity Recognition with Nearest Neighbor Search. arXiv:[2203.17103](https://arxiv.org/abs/2203.17103) [cs.CL] <https://arxiv.org/abs/2203.17103>
- [313] Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024. Self-Taught Evaluators. arXiv:[2408.02666](https://arxiv.org/abs/2408.02666) [cs.CL] <https://arxiv.org/abs/2408.02666>
- [314] Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O’Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Shepherd: A Critic for Language Model Generation. arXiv:[2308.04592](https://arxiv.org/abs/2308.04592) [cs.CL] <https://arxiv.org/abs/2308.04592>
- [315] Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhui Wang. 2025. VisualPRM: An Effective Process Reward Model for Multimodal Reasoning. arXiv:[2503.10291](https://arxiv.org/abs/2503.10291) [cs.CV] <https://arxiv.org/abs/2503.10291>
- [316] Xiayao Wang, Linfeng Song, Ye Tian, Dian Yu, Baolin Peng, Haitao Mi, Furong Huang, and Dong Yu. 2024. Towards Self-Improvement of LLMs via MCTS: Leveraging Stepwise Knowledge with Curriculum Preference Learning. arXiv:[2410.06508](https://arxiv.org/abs/2410.06508) [cs.LG] <https://arxiv.org/abs/2410.06508>
- [317] Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and Improve Robustness in NLP Models: A Survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 4569–4586. doi:[10.18653/v1/2022.nacl-main.339](https://doi.org/10.18653/v1/2022.nacl-main.339)
- [318] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=1PL1NIMMrw>
- [319] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023. Large Language Models Are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=BGvkWZEGt7>
- [320] Yue Wang, Qiužhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhusong Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Thoughts Are All Over the Place: On the Underthinking of o1-Like LLMs. arXiv:[2501.18585](https://arxiv.org/abs/2501.18585) [cs.CL] <https://arxiv.org/abs/2501.18585>
- [321] Yan Wang, Dongyang Ma, and Deng Cai. 2024. With Greater Text Comes Greater Necessity: Inference-Time Training Helps Long Text Generation. In *First Conference on Language Modeling*. <https://openreview.net/forum?id=dj9x6JuiD5>
- [322] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. <https://openreview.net/forum?id=y10DM6R2r3>
- [323] Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=5Nn2BLV7SB>
- [324] Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. 2024. RAT: Retrieval Augmented Thoughts Elicit Context-Aware Reasoning in Long-Horizon Generation. arXiv:[2403.05313](https://arxiv.org/abs/2403.05313) [cs.CL] <https://arxiv.org/abs/2403.05313>
- [325] Anjiang Wei, Jiannan Cao, Ran Li, Hongyu Chen, Yuhui Zhang, Ziheng Wang, Yuan Liu, Thiago S. F. X. Teixeira, Diyi Yang, Ke Wang, and Alex Aiken. 2025. EquiBench: Benchmarking Large Language Models’ Understanding of Program Semantics via Equivalence Checking. arXiv:[2502.12466](https://arxiv.org/abs/2502.12466) [cs.LG] <https://arxiv.org/abs/2502.12466>

- [326] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=_VjQLMeSB_J
- [327] Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating Sequences by Learning to Self-Correct. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=hH36JeQZDaO>
- [328] Jason Weston and Sainbayar Sukhbaatar. 2023. System 2 Attention (is something you might need too). arXiv:2311.11829 [cs.CL] <https://arxiv.org/abs/2311.11829>
- [329] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*. PMLR, 23965–23998.
- [330] Han Wu, Yuxuan Yao, Shuqi Liu, Zehua Liu, Xiaojin Fu, Xiongwei Han, Xing Li, Hui-Ling Zhen, Tao Zhong, and Mingxuan Yuan. 2025. Unlocking Efficient Long-to-Short LLM Reasoning with Model Merging. arXiv:2503.20641 [cs.CL] <https://arxiv.org/abs/2503.20641>
- [331] Jinyang Wu, Mingkuan Feng, Shuai Zhang, Feihu Che, Zengqi Wen, and Jianhua Tao. 2024. Beyond Examples: High-level Automated Reasoning Paradigm in In-Context Learning via MCTS. arXiv:2411.18478 [cs.CL] <https://arxiv.org/abs/2411.18478>
- [332] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind’s Eye of LLMs: Visualization-of-Thought Elicits Spatial Reasoning in Large Language Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=CEJ1mYPgWw>
- [333] Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024. Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for Problem-Solving with Language Models. arXiv:2408.00724 [cs.AI] <https://arxiv.org/abs/2408.00724>
- [334] Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1423–1436. doi:10.18653/v1/2023.acl-long.79
- [335] Zhiheng Xi, Dingwen Yang, Jixuan Huang, Jiafu Tang, Guanyu Li, Yiwen Ding, Wei He, Boyang Hong, Shihan Do, Wenyu Zhan, Xiao Wang, Rui Zheng, Tao Ji, Xiaowei Shi, Yitao Zhai, Rongxiang Weng, Jingang Wang, Xunliang Cai, Tao Gui, Zuxuan Wu, Qi Zhang, Xipeng Qiu, Xuanjing Huang, and Yu-Gang Jiang. 2024. Enhancing LLM Reasoning via Critique Models with Test-Time and Training-Time Supervision. arXiv:2411.16579 [cs.CL] <https://arxiv.org/abs/2411.16579>
- [336] Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. 2025. TokenSkip: Controllable Chain-of-Thought Compression in LLMs. arXiv:2502.12067 [cs.CL] <https://arxiv.org/abs/2502.12067>
- [337] Heming Xia, Zhe Yang, Qingxin Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, and Zhifang Sui. 2024. Unlocking efficiency in large language model inference: A comprehensive survey of speculative decoding. *arXiv preprint arXiv:2401.07851* (2024).
- [338] Wenyi Xiao, Leilei Gan, Weilong Dai, Wanggui He, Ziwei Huang, Haoyuan Li, Fangxun Shu, Zhelun Yu, Peng Zhang, Hao Jiang, and Fei Wu. 2025. Fast-Slow Thinking for Large Vision-Language Model Reasoning. arXiv:2504.18458 [cs.CL] <https://arxiv.org/abs/2504.18458>
- [339] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. 2025. SANA 1.5: Efficient Scaling of Training-Time and Inference-Time Compute in Linear Diffusion Transformer. arXiv:2501.18427 [cs.CV] <https://arxiv.org/abs/2501.18427>
- [340] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2025. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=o6Ynz6OIQ6>
- [341] Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P. Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte Carlo Tree Search Boosts Reasoning via Iterative Preference Learning. arXiv:2405.00451 [cs.AI] <https://arxiv.org/abs/2405.00451>
- [342] Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. 2023. Self-Evaluation Guided Beam Search for Reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=Bw82hwg5Q3>
- [343] Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-Reasoner: Improving Reasoning Capability in Large Audio Language Models. arXiv:2503.02318 [cs.SD] <https://arxiv.org/abs/2503.02318>
- [344] Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining Inter-Consistency of Large Language Models Collaboration: An In-depth Analysis via Debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 7572–7590. doi:10.18653/v1/2023.findings-emnlp.508
- [345] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712* (2024).
- [346] Bin Xu, Yiguan Lin, Yinghao Li, and Yang Gao. 2024. SRA-MCTS: Self-driven Reasoning Augmentation with Monte Carlo Tree Search for Code Generation. arXiv:2411.11053 [cs.CL] <https://arxiv.org/abs/2411.11053>
- [347] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. arXiv:2411.10440 [cs.CV] <https://arxiv.org/abs/2411.10440>
- [348] Haotian Xu. 2023. No Train Still Gain. Unleash Mathematical Reasoning of Large Language Models with Monte Carlo Tree Search Guided by Energy Function. arXiv:2309.03224 [cs.AI] <https://arxiv.org/abs/2309.03224>

- [349] Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025. Chain of Draft: Thinking Faster by Writing Less. arXiv:2502.18600 [cs.CL] <https://arxiv.org/abs/2502.18600>
- [350] Yuhui Xu, Hanze Dong, Lei Wang, Doyen Sahoo, Junnan Li, and Caiming Xiong. 2025. Scalable Chain of Thoughts via Elastic Reasoning. arXiv:2505.05315 [cs.LG] <https://arxiv.org/abs/2505.05315>
- [351] Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025. SoftCoT: Soft Chain-of-Thought for Efficient Reasoning with LLMs. arXiv:2502.12134 [cs.CL] <https://arxiv.org/abs/2502.12134>
- [352] Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025. SoftCoT++: Test-Time Scaling with Soft Chain-of-Thought Reasoning. arXiv:2505.11484 [cs.CL] <https://arxiv.org/abs/2505.11484>
- [353] Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Zheng Lin, Li Cao, and Weiping Wang. 2025. Dynamic Early Exit in Reasoning Models. arXiv:2504.15895 [cs.CL] <https://arxiv.org/abs/2504.15895>
- [354] Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. 2024. Regularizing Hidden States Enables Learning Generalizable Reward Model for LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=jwh9MHEfmY>
- [355] Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. 2025. Towards Thinking-Optimal Scaling of Test-Time Compute for LLM Reasoning. arXiv:2502.18080 [cs.CL] <https://arxiv.org/abs/2502.18080>
- [356] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazhen Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. 2025. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=LQzN6TRFg9>
- [357] Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, and Zhifang Sui. 2024. Confidence v.s. Critique: A Decomposition of Self-Correction Capability for LLMs. arXiv:2412.19513 [cs.CL] <https://arxiv.org/abs/2412.19513>
- [358] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and Dacheng Tao. 2024. Mulberry: Empowering MLLM with o1-like Reasoning and Reflection via Collective Monte Carlo Tree Search. arXiv:2412.18319 [cs.CV] <https://arxiv.org/abs/2412.18319>
- [359] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 11809–11822. https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf
- [360] Michihiro Yasunaga and Percy Liang. 2020. Graph-based, Self-Supervised Program Repair from Diagnostic Feedback. In *International Conference on Machine Learning (ICML)*.
- [361] Hai Ye, Yuyang Ding, Juntao Li, and Hwee Tou Ng. 2022. Robust Question Answering against Distribution Shifts with Test-Time Adaption: An Empirical Study. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 6179–6192. doi:[10.18653/v1/2022.findings-emnlp.460](https://doi.org/10.18653/v1/2022.findings-emnlp.460)
- [362] Hai Ye and Hwee Tou Ng. 2024. Preference-Guided Reflective Sampling for Aligning Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 21646–21668. doi:[10.18653/v1/2024.emnlp-main.1206](https://doi.org/10.18653/v1/2024.emnlp-main.1206)
- [363] Hai Ye, Qizhe Xie, and Hwee Tou Ng. 2023. Multi-Source Test-Time Adaptation as Dueling Bandits for Extractive Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 9647–9660. doi:[10.18653/v1/2023.acl-long.537](https://doi.org/10.18653/v1/2023.acl-long.537)
- [364] Zihuiwen Ye, Fraser Greenlee-Scott, Max Bartolo, Phil Blunsom, Jon Ander Campos, and Matthias Gallé. 2024. Improving Reward Models with Synthetic Critiques. arXiv:2405.20850 [cs.CL] <https://arxiv.org/abs/2405.20850>
- [365] Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying Long Chain-of-Thought Reasoning in LLMs. arXiv:2502.03373 [cs.CL] <https://arxiv.org/abs/2502.03373>
- [366] Jingyang Yi, Jiaheng Wang, and Sida Li. 2025. ShorterBetter: Guiding Reasoning Models to Find Optimal Inference Length for Efficient Reasoning. arXiv:2504.21370 [cs.AI] <https://arxiv.org/abs/2504.21370>
- [367] Wangjie You, Pei Guo, Juntao Li, Kehai Chen, and Min Zhang. 2024. Efficient Domain Adaptation for Non-Autoregressive Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 13657–13670. doi:[10.18653/v1/2024.findings-acl.810](https://doi.org/10.18653/v1/2024.findings-acl.810)
- [368] Fei Yu, Yingru Li, and Benyou Wang. 2025. Scaling Flaws of Verifier-Guided Search in Mathematical Reasoning. arXiv:2502.00271 [cs.CL] <https://arxiv.org/abs/2502.00271>
- [369] Hao Yu, Bo Shen, Dezhui Ran, Jiaxin Zhang, Qi Zhang, Yuchi Ma, Guangtai Liang, Ying Li, Qianxiang Wang, and Tao Xie. 2024. Codereval: A benchmark of pragmatic code generation with generative pre-trained models. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*. 1–12.
- [370] Jiachen Yu, Shaoning Sun, Xiaohui Hu, Jiaxu Yan, Kaidong Yu, and Xuelong Li. 2025. Improve LLM-as-a-Judge Ability as a General Ability. arXiv:2502.11689 [cs.CL] <https://arxiv.org/abs/2502.11689>
- [371] Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. 2024. Distilling System 2 into System 1. arXiv:2407.06023 [cs.CL] <https://arxiv.org/abs/2407.06023>
- [372] Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, Melanie Kambadur, Dhruv Mahajan, and Rui Hou. 2024. Self-Generated Critiques Boost Reward Modeling for Language Models.

- arXiv:2411.16646 [cs.CL] <https://arxiv.org/abs/2411.16646>
- [373] Yongcan Yu, Lijun Sheng, Ran He, and Jian Liang. 2023. Benchmarking test-time adaptation against distribution shifts in image classification. *arXiv preprint arXiv:2307.03133* (2023).
- [374] Danlong Yuan, Tian Xie, Shaohan Huang, Zhuocheng Gong, Huishuai Zhang, Chong Luo, Furu Wei, and Dongyan Zhao. 2025. Efficient RL Training for Reasoning Models via Length-Aware Optimization. arXiv:2505.12284 [cs.AI] <https://arxiv.org/abs/2505.12284>
- [375] Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. 2024. Advancing LLM Reasoning Generalists with Preference Trees. arXiv:2404.02078 [cs.AI] <https://arxiv.org/abs/2404.02078>
- [376] Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. 2024. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981* (2024).
- [377] Shuozi Yuan, Liming Chen, Miaomiao Yuan, Jin Zhao, Haoran Peng, and Wenming Guo. 2025. MCTS-SQL: An Effective Framework for Text-to-SQL with Monte Carlo Tree Search. arXiv:2501.16607 [cs.DB] <https://arxiv.org/abs/2501.16607>
- [378] Siyu Yuan, Zehui Chen, Zhiheng Xi, Junjie Ye, Zhengyin Du, and Jiecao Chen. 2025. Agent-R: Training Language Model Agents to Reflect via Iterative Self-Training. arXiv:2501.11425 [cs.AI] <https://arxiv.org/abs/2501.11425>
- [379] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-Rewarding Language Models. arXiv:2401.10020 [cs.CL] <https://arxiv.org/abs/2401.10020>
- [380] Parvez Zamil and Gollam Rabby. 2024. AIME Problems 1983 to 2024. doi:10.34740/KAGGLE/DSV/8834060
- [381] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STaR: Bootstrapping Reasoning With Reasoning. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=_3ELRdg2sgI
- [382] Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, Rongwu Xu, Zehan Qi, Wanru Zhao, et al. 2024. MR-BEN: A Comprehensive Meta-Reasoning Benchmark for Large Language Models. *arXiv preprint arXiv:2406.13975* (2024).
- [383] Yuanzhao Zhai, Tingkai Yang, Kele Xu, Feng Dawei, Cheng Yang, Bo Ding, and Huaimin Wang. 2024. Enhancing Decision-Making for LLM Agents via Step-Level Q-Value Models. arXiv:2409.09345 [cs.AI] <https://arxiv.org/abs/2409.09345>
- [384] Runzhe Zhan, Xuebo Liu, Derek F. Wong, Cuilian Zhang, Lidia S. Chao, and Min Zhang. 2023. Test-time Adaptation for Machine Translation Evaluation by Uncertainty Minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 807–820. doi:10.18653/v1/2023.acl-long.47
- [385] Di Zhang, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. 2024. Accessing GPT-4 level Mathematical Olympiad Solutions via Monte Carlo Tree Self-refine with LLaMa-3 8B. arXiv:2406.07394 [cs.AI] <https://arxiv.org/abs/2406.07394>
- [386] Di Zhang, Jiabao Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, Wanli Ouyang, and Dongzhan Zhou. 2024. LLaMA-Berry: Pairwise Optimization for O1-like Olympiad-Level Mathematical Reasoning. arXiv:2410.02884 [cs.AI] <https://arxiv.org/abs/2410.02884>
- [387] Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. ReST-MCTS*: LLM Self-Training via Process Reward Guided Tree Search. arXiv:2406.03816 [cs.CL] <https://arxiv.org/abs/2406.03816>
- [388] Fengji Zhang, Bei Chen, Yue Zhang, Jacky Keung, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. 2023. RepoCoder: Repository-Level Code Completion Through Iterative Retrieval and Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2471–2484. doi:10.18653/v1/2023.emnlp-main.151
- [389] Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. 2024. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. *arXiv preprint arXiv:2410.02506* (2024).
- [390] Hanning Zhang, Pengcheng Wang, Shizhe Diao, Yong Lin, Rui Pan, Hanze Dong, Dylan Zhang, Pavlo Molchanov, and Tong Zhang. 2024. Entropy-Regularized Process Reward Model. *arXiv preprint arXiv:2412.11006* (2024).
- [391] Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. 2025. AdaptThink: Reasoning Models Can Learn When to Think. arXiv:2505.13417 [cs.CL] <https://arxiv.org/abs/2505.13417>
- [392] Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. 2025. LightThinker: Thinking Step-by-Step Compression. arXiv:2502.15589 [cs.CL] <https://arxiv.org/abs/2502.15589>
- [393] Kexun Zhang, Shang Zhou, Danqing Wang, William Yang Wang, and Lei Li. 2024. Scaling LLM Inference with Optimized Sample Compute Allocation. arXiv:2410.22480 [cs.CL] <https://arxiv.org/abs/2410.22480>
- [394] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024. Generative Verifiers: Reward Modeling as Next-Token Prediction. arXiv:2408.15240 [cs.LG] <https://arxiv.org/abs/2408.15240>
- [395] Marvin Zhang, Sergey Levine, and Chelsea Finn. 2022. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems* 35 (2022), 38629–38642.
- [396] Qingjie Zhang, Han Qiu, Di Wang, Haotong Qian, Yiming Li, Tianwei Zhang, and Minlie Huang. 2024. Understanding the Dark Side of LLMs' Intrinsic Self-Correction. arXiv:2412.14959 [cs.CL] <https://arxiv.org/abs/2412.14959>
- [397] Ruiqi Zhang, Momin Haider, Ming Yin, Jiahao Qiu, Mengdi Wang, Peter Bartlett, and Andrea Zanette. 2024. Accelerating Best-of-N via Speculative Rejection. In *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML)*

- 2024). <https://openreview.net/forum?id=dRp8tAIPhj>
- [398] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024. Improve Vision Language Model Chain-of-thought Reasoning. arXiv:2410.16198 [cs.AI] <https://arxiv.org/abs/2410.16198>
- [399] Wenyuan Zhang, Shuaiyi Nie, Xinghua Zhang, Zefeng Zhang, and Tingwen Liu. 2025. S1-Bench: A Simple Benchmark for Evaluating System 1 Thinking Capability of Large Reasoning Models. arXiv:2504.10368 [cs.CL] <https://arxiv.org/abs/2504.10368>
- [400] Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474* (2023).
- [401] Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active Example Selection for In-Context Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9134–9148. doi:10.18653/v1/2022.emnlp-main.622
- [402] Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2023. Adanpc: Exploring non-parametric classifier for test-time adaptation. In *International Conference on Machine Learning*. PMLR, 41647–41676.
- [403] Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. 2024. o1-Coder: an o1 Replication for Coding. arXiv:2412.00154 [cs.SE] <https://arxiv.org/abs/2412.00154>
- [404] Zhihan Zhang, Tao Ge, Zhenwen Liang, Wenhao Yu, Dian Yu, Mengzhao Jia, Dong Yu, and Meng Jiang. 2024. Learn Beyond The Answer: Training Language Models with Reflection for Mathematical Reasoning. arXiv:2406.12050 [cs.CL] <https://arxiv.org/abs/2406.12050>
- [405] Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. 2024. Multimodal Chain-of-Thought Reasoning in Language Models. *Transactions on Machine Learning Research* (2024). <https://openreview.net/forum?id=y1pPWVFvF>
- [406] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=5NTt8GFjUHkr>
- [407] Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. The Lessons of Developing Process Reward Models in Mathematical Reasoning. *arXiv preprint arXiv:2501.07301* (2025).
- [408] Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. 2023. On Pitfalls of Test-Time Adaptation. In *International Conference on Machine Learning*. PMLR, 42058–42080.
- [409] Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, and Bowen Zhou. 2025. GenPRM: Scaling Test-Time Compute of Process Reward Models via Generative Reasoning. arXiv:2504.00891 [cs.CL] <https://arxiv.org/abs/2504.00891>
- [410] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. 2024. Test-Time Adaptation with CLIP Reward for Zero-Shot Generalization in Vision-Language Models. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=kIP0duasBb>
- [411] Weixiang Zhao, Xingyu Sui, Jiahe Guo, Yulin Hu, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che, Tat-Seng Chua, and Ting Liu. 2025. Trade-offs in Large Reasoning Models: An Empirical Analysis of Deliberative and Adaptive Reasoning over Foundational Capabilities. arXiv:2503.17979 [cs.AI] <https://arxiv.org/abs/2503.17979>
- [412] Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards Open Reasoning Models for Open-Ended Solutions. arXiv:2411.14405 [cs.CL] <https://arxiv.org/abs/2411.14405>
- [413] Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559* (2024).
- [414] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. <https://openreview.net/forum?id=uccHPGDlao>
- [415] Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. DeepResearcher: Scaling Deep Research via Reinforcement Learning in Real-world Environments. arXiv:2504.03160 [cs.AI] <https://arxiv.org/abs/2504.03160>
- [416] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=WZH7099tgfM>
- [417] Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. RMB: Comprehensively Benchmarking Reward Models in LLM Alignment. arXiv:2410.09893 [cs.CL] <https://arxiv.org/abs/2410.09893>
- [418] Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. 2024. Image-of-Thought Prompting for Visual Reasoning Refinement in Multimodal Large Language Models. arXiv:2405.13872 [cs.AI] <https://arxiv.org/abs/2405.13872>
- [419] Zhi Zhou, Tan Yuhao, Zenan Li, Yuan Yao, Lan-Zhe Guo, Xiaoxing Ma, and Yu-Feng Li. 2025. Bridging Internal Probability and Self-Consistency for Effective and Efficient LLM Reasoning. arXiv:2502.00511 [cs.LG] <https://arxiv.org/abs/2502.00511>
- [420] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. JudgeLM: Fine-tuned Large Language Models are Scalable Judges. arXiv:2310.17631 [cs.CL] <https://arxiv.org/abs/2310.17631>
- [421] Wenhao Zhu, Shujian Huang, Yunzhe Lv, Xin Zheng, and Jiajun Chen. 2023. What Knowledge Is Needed? Towards Explainable Memory for kNN-MT Domain Adaptation. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 2824–2836. doi:10.18653/v1/2023.findings-acl.177

- [422] Yuhan Zhu, Guozhen Zhang, Chen Xu, Haocheng Shen, Xiaoxin Chen, Gangshan Wu, and Limin Wang. 2024. Efficient Test-Time Prompt Tuning for Vision-Language Models. arXiv:[2408.05775](https://arxiv.org/abs/2408.05775) [cs.CV] <https://arxiv.org/abs/2408.05775>
- [423] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proc. IEEE* 109, 1 (2020), 43–76.
- [424] Terry Yue Zhuo, Vu Minh Chien, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widayarsi, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, Simon Brunner, Chen GONG, James Hoang, Armel Randy Zebaze, Xiaoheng Hong, Wen-Ding Li, Jean Kaddour, Ming Xu, Zhihan Zhang, Prateek Yadav, Naman Jain, Alex Gu, Zhoujun Cheng, Jiawei Liu, Qian Liu, Zijian Wang, David Lo, Binyuan Hui, Niklas Muennighoff, Daniel Fried, Xiaoning Du, Harm de Vries, and Leandro Von Werra. 2025. BigCodeBench: Benchmarking Code Generation with Diverse Function Calls and Complex Instructions. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=YrycTjllL0>

Received 27 June 2025